International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 3, June 2021, pp. 2407~2413
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i3.pp2407-2413

Prediction model of algal blooms using logistic regression and confusion matrix

Hongwon Yun
Department of Computer Software Engineering, Silla University, Busan, Republic of Korea

Article Info

Article history:

Received Jul 31, 2020
Revised Sep 22, 2020
Accepted Oct 8, 2020

Keywords:

Algal blooms
Confusion matrix
Ensemble method
Logistic regression
Prediction model

ABSTRACT

Algal blooms data are collected and refined as experimental data for algal blooms prediction. Refined algal blooms dataset is analyzed by logistic regression analysis, and statistical tests and regularization are performed to find the marine environmental factors affecting algal blooms. The predicted value of algal bloom is obtained through logistic regression analysis using marine environment factors affecting algal blooms. The actual values and the predicted values of algal blooms dataset are applied to the confusion matrix. By improving the decision boundary of the existing logistic regression, and accuracy, sensitivity and precision for algal blooms prediction are improved. In this paper, the algal blooms prediction model is established by the ensemble method using logistic regression and confusion matrix. Algal blooms prediction is improved, and this is verified through big data analysis.

Corresponding Author:

Hongwon Yun
Department of Computer Software Engineering
Silla University
Busan 46958, Republic of Korea
Email: hwyun@silla.ac.kr

1. INTRODUCTION

Logistic regression is a special case of a typical model and is similar to linear regression, however it has a difference in the relationship between dependent and independent variables. The dependent variable of logistic regression can be binary or continuous, and it is used as a model for classification or prediction when the dependent variable is binary [1, 2]. If the dependent variable of logistic regression is binary, the range of its value is limited to the bivariate and the distribution of conditional probability follows the Bernoulli distribution. Logistic regression allows dependent variable values to be between 0 and 1 regardless of the range of independent variable values, so it is possible to classify the result of data into a specific classification when input data is given and predict the likelihood of an event occurring [3-5].

In logistic regression, where the dependent variable is binary, the predicted value can be calculated using a linear combination of the independent variables. However, since the value of the dependent variable is classified as pass or fail around the decision boundary, the value close to the decision boundary may be less accurate [6-8]. In binary logistic regression, since the actual value of the dependent variable is present and the predicted value can be calculated, the predicted value can be applied to a confusion matrix that can be compared to the target value [9, 10]. It can be obtained sensitivity and precision from the confusion matrix using the actual and predicted values of the logistic regression, and apply it to algal blooms to create a summary of indicators such as sensitivity and precision including accuracy [11-13].

Sensitivity and precision are as important as accuracy in predicting algal bloom occurrence. This is because high sensitivity and precision can provide indicators that can prevent massive property damage [14-17]. The elements of the marine environment that cause algal blooms are generally known, but no study

can be found to analyze the influence of each element on algal blooms and predict algal blooms. In this study, the predicted value of logistic regression is calculated by machine learning. The actual value used in logistic regression analysis and the prediction calculated through machine running are applied to the confusion matrix to create a prediction model for algal blooms.

This paper is organized as follows. The logistic regression and confusion matrix as the background theory of this study are describe in section 2. In section 3, we describe the algal blooms prediction model using the ensemble method of the logistic regression and confusion matrix proposed in this study. Here we describe the process of extracting marine environmental elements using logistic regression, obtaining red tide prediction values, applying improved decision boundaries to logistic regression, and how to improve accuracy, sensitivity and precision through confusion matrix. In section 4, we verify the proposed algal blooms prediction model using the algal blooms dataset, and conclusions are described in section 5.

2. LOGISTIC REGRESSION AND CONFUTION MATRIX

2.1. Logistic regression

Linear regression is a model that estimates a regression coefficient that can linearly express the relationship between independent variables *X* and dependent variables *Y* with continuous values. If the dependent variable *Y* is a binary variable, logistic regression is used because linear regression cannot be applied directly. Some regression algorithms can be used for classification, and logistic regression is widely used to estimate the probability that a sample belongs to a particular class. If the estimated probability exceeds 0.5, the sample is predicted to belong to the class, and if it is less than 0.5, it is used as a binary classifier to predict that the sample does not belong to the class [18, 19]. To estimate the probability, logistic regression calculates the weighted sum of the input characteristics, but instead of outputting the result immediately such as linear regression, it outputs the logistic of the result value. Logistic is a sigmoid function that outputs a value between 0 and 1 [20]. The logistic function has the effect of limiting the output result to always between 0 and 1 for numerical values *x*, and its expression is defined as follows.

$$y = \frac{1}{1+e^{-f(x)}} \tag{1}$$

In (1), $f(x)$ can be either a simple linear function or a multiple linear function. For classification problems with two categories, if $f(x) > 0$ is classified as $y \to 1$ and $f(x) < 0$ is classified as $y \to 0$. The decision boundary of the logistic regression model is the $f(x) = 0$ in hyperplane and becomes $y = 0.5$. Errors in prediction usually occur around the decision boundary [21, 22].

2.2. Confusion matrix

The confusion matrix is a tool that easily and effectively shows the performance of the classifier and has the advantage of being easy to interpret the results. A confusion matrix can be used to evaluate the performance of any models or algorithms. As shown in Table 1, the rows in the confusion matrix represent the values of the predictive class and the columns represent the values of the actual class. Each cell is one of the possible combinations of prediction and actuality. In the 2×2 confusion matrix, there are true positive (TP), false positive (FP), false negative (FN), and true false (TF) [23].

The perfect model will only have values on the diagonal, the rest of the cells will be all zeros, and the bad model will be evenly distributed in all cells. The error matrix tells us how bad a model is when it is bad. The value of each cell can identify a misclassified pattern [24].

Table 1. Confusion matrix

Confusion matrix		True class (Actual)	
		P	N
Hypothesized class (Predicted)	Y	True Positives	False Positives
	N	False Negatives	True Negatives

Methods for summarizing the results of the confusion matrix include accuracy, precision, and recall.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

The accuracy is obtained by dividing the accurately predicted number (*TP*+*TN*) by the total number of samples, and is represented by (2). Among the methods for summarizing the results in the confusion

matrix, the most frequently used precision and sensitivity are as shown in (3) and (4), respectively.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Precision is a positive predictive value that measures how many of the samples (*TP+FP*) predicted to be positive are true positives (*TP*). Precision is used as a performance indicator when the goal is to reduce the number of false positives (*FP*).

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

Sensitivity measures how many of the total positive samples (*TP+FN*) are classified as positive classes (*TP*).

3. PREDICTION MODEL

After collecting algal blooms dataset from the National Institute of Fisheries Science, it was cleaned and refined. The first multiple logistic regression analysis was performed on the refined algal blooms dataset, and some attributes were removed through a statistical test. A second multiple logistic regression analysis was performed with the exception of the attributes removed and then the regularization was applied. After applying the regularization, a third multiple logistic regression analysis is performed and the results are applied to the confusion matrix. Figure 1 shows this process.

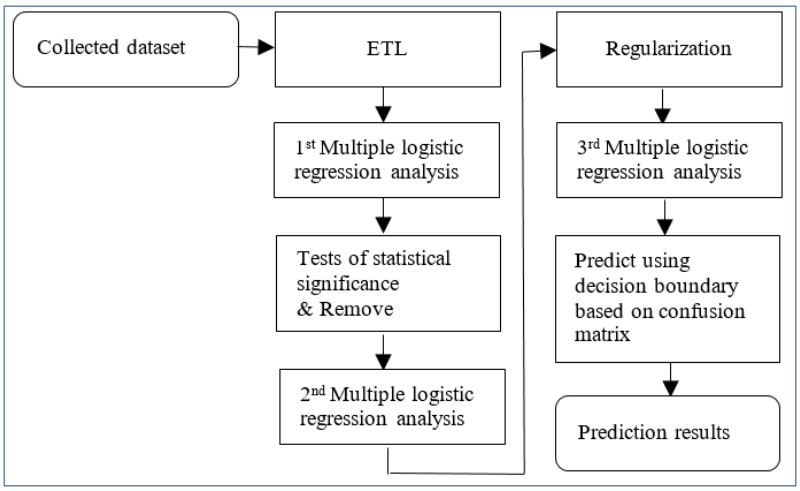


Figure 1. Prediction process of algal blooms

The probability of occurrence of harmful algal blooms with two or more independent variables is defined as $p(x)$ and the odds as $= \frac{p}{1-p}$. When the range of input values is [0, 1], log it transformation is performed to adjust the range of output values to $(-\infty, \infty)$, resulting in $\log(odds) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$ [25]. Therefore, for multiple independent variables that affect harmful algal blooms, the multiple logistic function that allows the dependent variable range to be between [0, 1] is as shown in (5). In (5) calculates the effect of each element of the ocean observation data, which is an independent variable, on the occurrence of a harmful algal blooms as a dependent variable. This is a basic model for estimating the probability of occurrence of harmful algal blooms.

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i)}} \quad (5)$$

The maximum likelihood estimation is used to estimate parameter β in regression expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$ by logit transformation. The log likelihood function can be obtained from the likelihood function [26] expressed as the product of Bernoulli's probability function, and is expressed as (6).

The parameter that maximizes the log likelihood function in (6) is determined from multiple independent variables that affect the harmful algal blooms.

$$\ln L = \sum_i y_i(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i) + \sum_i \ln\left(1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i}\right) \quad (6)$$

The L1 regularization [27] used to eliminate low-impact independent variables among multiple independent variables that affect harmful algal blooms is shown in (7).

$$\text{argmin} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (7)$$

The properties in the marine environment observation dataset are shown in Table 2 and used as independent variables in logistic regression.

Table 2. Multiple independent variables for logistic regression

Variables	Comments
T	Temperature
S	Salinity
DO	Dissolved Oxygen
P	Phosphate Phosphorus
NA	Nitrous Acid Nitrogen
N	Nitric Acid Nitrogen
SA	Silicic Acid Silicon

In (8) is obtained by applying seven independent variables, such as water temperature, salinity, dissolved oxygen, phosphate phosphorus, nitrous acid nitrogen, nitric acid nitrogen, silicic acid silicon, to the basic model of multiple logistic regression (5).

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 DO_i + \beta_4 P_i + \beta_5 NA_i + \beta_6 N_i + \beta_7 SA_i \quad (8)$$

P-value is used to determine if any independent variable was statistically significant in the results of multiple logistic regression analysis on the training dataset, and independent variables with a P-value of 0.05 or higher are excluded. The parameters for statistically significant independent variables are as shown in (9).

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 P_i + \beta_4 N_i \quad (9)$$

The regulation for removing an independent variable close to zero in order to make some coefficients zero is as shown in (7). The result is as shown in (10) when (7) is applied to the result of (9).

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 P_i \quad (10)$$

In (11) is the logistic regression model for algal blooms prediction obtained by applying the above process to the algal blooms dataset.

$$p(x) = \frac{1}{1+e^{-(-5.89+0.34T_i-0.12S_i+0.35P_i)}} \quad (11)$$

The normalization process from the (8) to the (11) is from Step 2 to Step 5 among the algorithms in Table 3, respectively. The algal blooms prediction model was normalized while performing experiments based on the algorithm in Table 3. The detailed experimental process is described in section 4. The equation for obtaining a decision boundary to increase the sensitivity and precision is defined as shown in (12).

$$\Delta = \left|0.5 \pm \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP}\right)\right| \quad (12)$$

Table 3 shows algorithm for establishing algal blooms prediction model. This algorithm shows the process of performing multiple logistic regression first in a refined dataset, then statistical tests on the results,

and then removing low-weight independent variables, finally setting up a logistic regression model, and finding the decision boundary finally.

Table 3. Algorithm for establishing algal blooms prediction model

Step	Statements
1	Extraction, Transformation and Loading from collected dataset Prepare training dataset
2	Perform multiple regression analysis using (1) on the training dataset Output regression coefficients and statistical tests
3	Perform a statistical significance test ·attributes P-value > 0.5 are excluded in the training dataset
4	Perform multiple regression analysis for the training dataset with attributes whose P-value <= 0.5 Output regression coefficients and statistical tests for attributes whose P-value <= 0.5
5	Regularize regression coefficients from step 4 using $\text{argmin} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$ Perform multiple regression analysis using a regularized regression formula Output test dataset
6	Input test dataset from step 5 Predict probability using decision boundary $\Delta = \left\lvert 0.5 \pm \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP}\right)\right\rvert$based on confusion matrix

4. EXPERIMENT

Multiple logistic regression analysis (8) can be performed on the training dataset to obtain the results shown in Table 4. In Table 4, p-value is used to determine whether any independent variable is statistically significant, and independent variables with a p-value of 0.05 or higher are excluded. Parameters β are determined for statistically significant independent variables in Table 4 and L1 regularization is applied and then the results shown in Table 5 can be obtained. In (11) of the logistic regression model for algal blooms prediction is obtained from the coefficients in Table 5.

Table 4. 1st multiple logistic regression analysis on training dataset

Input variables	Coefficient	Std. error	P-value
Constant	-5.35	1.49	0.00
Temperature	0.33	0.02	0.00
Salinity	-0.12	0.03	0.00
Dissolved Oxygen	-0.05	0.08	0.54
Phosphate Phosphorus	0.38	0.14	0.01
Nitrous Acid Nitrogen	-0.07	0.12	0.58
Nitric Acid Nitrogen	-0.06	0.02	0.00
Silicic Acid Silicon	-0.02	0.01	0.16

Table 5. Coefficients of logistic regression model for algal blooms prediction

Input variables	Coefficient	Std. error	P-value
Constant	-5.89	1.21	0.00
Temperature	0.34	0.01	0.00
Salinity	-0.12	0.03	0.00
Phosphate Phosphorus	0.35	0.14	0.01

Predicting the occurrence of algal blooms from (11) gives 91.84% accuracy. Accuracy alone may not be sufficient to assess the predicted performance of algal blooms. We utilize the confusion matrix since we do not know false negatives or false positives of algal blooms. A confusion matrix for algal blooms shown in Table 6 is obtained from algal blooms dataset.

Table 6. Confusion matrix for algal blooms

Confusion matrix (Error matrix)		Actual values of algal blooms P(Occurrence)	N(Not occurrence)
Predicted values of algal blooms	Y (0.5 or higher)	True Positive tp=222	False Positive fp=205
	N (less than 0.5)	False Negative fn=599	True Negative tn=8828

Table 7 shows the sensitivity, specificity, and precision are obtained based on the decision boundary 0.5 using the values of the confusion matrix in Table 6. The prediction rate of false negative of algal blooms is low as the sensitivity is 27.04%, and the prediction rate of false positive is also low because the precision is 51.99%. Since the sensitivity and precision are low in case of the decision boundary is 0.5, we apply proposed the decision boundary (12) in order to solve these problems, and results are as shown in Table 8. When the decision boundary proposed in this paper is applied, the decision boundary becomes $\Delta = |0.5 \pm 0.25|$. When this is used as a decision boundary, TP=494, TN=9026, FN=327, FP=7, the sensitivity is 60.17%, and the precision is 98.6% as shown in Table 8.

Table 7. Resulting confusion matrix based on decision boundary 0.5 (unit: %)

TPR	TNR	PPV	FPR	ACC	F1
Sensitivity	Specificity	Precision	Fallout	Accuracy	F1 Score
27.04	97.73	51.99	2.27	91.84	35.58

Table 8. Resulting confusion matrix based on proposed decision boundary Δ (unit: %)

TPR	TNR	PPV	FPR	ACC	F1
Sensitivity	Specificity	Precision	Fallout	Accuracy	F1 Score
60.17	99.92	98.6	0.08	96.61	74.74

5. CONCLUSION

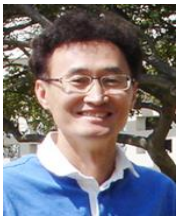
In this paper, logistic regression and confusion matrix were used to predict the occurrence of algal blooms. Algal blooms datasets were collected and refined for experimental analysis of algal blooms prediction. Logistic regression analysis was performed on refined algal blooms dataset and main marine environmental factors affecting algal blooms were found through statistical test and regularization processes. Logistic regression was performed using the marine environmental factors that were influential on algal blooms and the accuracy of algal bloom occurrence was obtained. The values of the confustion matrix were obtained using the dataset for algal blooms prediction and the predicted values obtained from logistic regression. Although the sensitivity and precision for the occurrence of algal blooms can be obtained from the values of the confusion matrix, the sensitivity and precision were low when the existing decision boundary was 0.5. Sensitivity and precision were improved by using the decision boundary proposed in this study. In this paper, the algal blooms prediction model was established by the ensemble method using logistic regression analysis and confusion matrix. Also, the accuracy, sensitivity, and precision for algal blooms prediction were improved, and these were verified through big data analysis.

REFERENCES

[1] Hosmer Jr., et al., "Applied logistic regression," *John Wiley & Sons*, vol. 398, 2013.
[2] M. Chang, et al., "Selection of Transformations of Continuous Predictors in Logistic Regression," *Information Technology-New Generations*, pp. 443-447, 2018.
[3] J. W. Osborne, "Simple Linear Models with Categorical Dependent Variables: Binary Logistic Regression," *SAGE*, pp. 97-132, 2017.
[4] J. Tolles and J. M. William, "Logistic regression: relating patient characteristics to outcomes," *Jama*, vol. 316, no. 5, pp. 533-534, 2016.
[5] A. D. Caigny, et al., "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *European Journal of Operational Research*, vol. 269, no. 2, pp. 760-772, 2018.
[6] X. Wan, "The Influence of Polynomial Order in Logistic Regression on Decision Boundary," *IOP Conference Series: Earth and Environmental Science*, vol. 267, no. 4, pp. 1-4, 2019.
[7] S. Ghazaal and A. Hakan, "Using the Distance in Logistic Regression Models for Predictor Ranking in Diabetes Detection," *International Conference on Medical and Biological Engineering*, 2019, pp. 665-670.
[8] J. Friedman, et al., "Additive logistic regression: A statistical view of boosting," *Annals of statistics*, vol. 28, no. 2, pp. 337-374, 2000.
[9] H. M. Ramos, et al., "A new explanatory index for evaluating the binary logistic regression based on the sensitivity of the estimated model," *Statistics and Probability Letters*, vol. 120, pp. 135-140, 2017.
[10] M. Ohsaki, et al., "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1806-1819, 2017.
[11] J. A. McGowan, et al., "Predicting coastal algal blooms in southern California," *Ecology*, vol. 98, no. 5, pp. 1419-1433, 2017.
[12] N. F. Manning, et al., "Extending the forecast model: Predicting Western Lake Erie harmful algal blooms at multiple spatial scales," *Journal of Great Lakes Research*, vol. 45, no. 3 pp. 587-595, 2019.

[13] N. Mellios, et al., "Machine Learning Approaches for Predicting Health Risk of Cyanobacterial Blooms in Northern European Lakes," *Water*, vol. 12, no. 4, p. 1191, 2020.
[14] L. Wang, et al., "An approach of improved Multivariate Timing-Random Deep Belief Net modelling for algal bloom prediction," *Biosystems engineering*, vol. 177, pp. 130-138, 2019.
[15] Ghatkar, et al., "Classification of algal bloom species from remote sensing data using an extreme gradient boosted decision tree model," *International Journal of Remote Sensing*, vol. 40, no. 24, pp. 9412-9438, 2019.
[16] S. Lee and D. Lee, "Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models," *International journal of environmental research and public health*, vol. 15, no. 7, p. 1322, 2018.
[17] X. Sun, et al., "A Bayesian structural model for predicting algal blooms," *Journal of Forecasting*, vol. 38, no. 8, pp. 788-802, 2019.
[18] F. Thabtah et al., "A machine learning autism classification based on logistic regression analysis," *Health information science and systems*, vol. 7, no. 1, p. 12, 2019.
[19] D. Menezes, et al., "Data classification with binary response through the Boosting algorithm and logistic regression," *Expert Systems with Applications*, vol. 69, pp. 62-73, 2017.
[20] J. M. Hilbe, "Practical guide to logistic regression," *CRC Press*, 2016.
[21] C. Fernández and F. Provost, "Causal Classification: Treatment Effect vs. Outcome Prediction," *Outcome Prediction*, 2019.
[22] K. Lee, et al., "Unbalanced data, type II error, and nonlinearity in predicting M&A failure," *Journal of Business Research*, vol. 109, pp. 271-287, 2010.
[23] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77-89, 1997.
[24] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.
[25] J. S. Cramer, "The origins and development of the logit model," *Cambridge UP*, pp. 1-19, 2003.
[26] I. M. Myung, "Tutorial on Maximum Likelihood Estimation," *Journal of Mathematical Psychology*, vol. 47, no. 1, pp. 90-100, 2003.
[27] F. Santosa and W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM Journal on Scientific and Statistical Computing*, *SIAM*, vol. 7, no. 4, pp. 1307-1330, 1986.

BIOGRAPHY OF AUTHOR



Hongwon Yun is a Professor with the Department of Computer Software Engineering at Silla University, Busan, South Korea. He received his B.S. and the Ph.D. degrees at the Department of Computer Science from Pusan National University, South Korea. His research interests include database system, big data analysis, and machine learning.