Marine Debris Prediction Based on Machine Learning Mrthods

楊舒晴
Team 25, Student ID: 106091221
Department of Education and Learning Technology, National Tsing Hua University,
Hsinchu, Taiwan

張浚騰
Team 25, Student ID: 106091228
Interdisciplinary Program of Electrical Engineering and Computer Science,
National Tsing Hua University,
Hsinchu, Taiwan

張峰榮
Team 25, Student ID: 109061534
Department of Electrical Engineering,
National Tsing Hua University,
Hsinchu, Taiwan

***Abstract*—In this work, we propose several machine learning methods based on the station information provided by SOW to predict the pollution level of station and compare the results of these methods. We expect the result can be use to reduce the need of station and manpower.**

***Keywords*—data preprocessing, regression, neuron network, random forest, Adaboost, Gradient-Boosting**

I. Introduction

Disposal of marine debris has long been a major issue in environment. It requires spending a lot of manpower and money. Therefore, how to utilize resource efficiently become a bottleneck. If we could predict the pollution level of the region, we could plan and make proper arrangement in advance. In this project, we tried to build up an Artificial Neural Network and to compare the results with other models such as adaboost, regression, gradient-boosting, and regression to solve the problem. Here is our flow chart about the project.

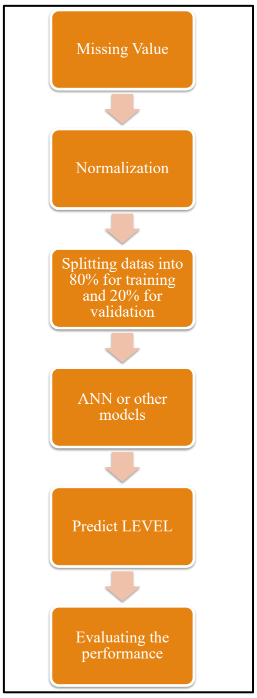

Missing Value
Normalization
Splitting datas into 80% for training and 20% for validation
ANN or other models
Predict LEVEL
Evaluating the performance

figure 1 Simple flowchart of our project

II. Related Work

Adaboost. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but if the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner. Coming to the advantages, Adaboost is less prone to overfitting as the input parameters are not jointly optimized. The accuracy of weak classifiers can be improved by using Adaboost. Nowadays, Adaboost is being used to classify text and images rather than binary classification problems.

The main disadvantage of Adaboost is that it needs a quality dataset. Noisy data and outliers have to be avoided before adopting an Adaboost algorithm.

Random forest. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Gradient-Boosting. Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

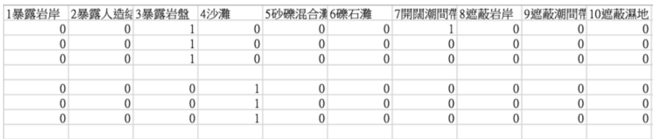
Artificial Neuron Network. Artificial neural networks are the modeling of the human brain with the simplest definition and building blocks are neurons. ANN are easy to conceptualize, and the learning methods are quite robust to noise in the training data because, the training examples may contain errors, which do not affect the final output. However, it also suffered from difficulty of showing the problem to the network, there is no specific rule for determining the structure of artificial neural networks, so the appropriate network structure is achieved through experience and trial and error which is time-consuming. [1]

III. Methods

A. *Data Preprocessing*

Before we train the model. We need to do data preprocessing which is crucial for the performance of our model.

Missing data. The structure contains following features:



1暴露岩岸	2暴露人造結	3暴露岩盤	4沙灘	5砂礫混合灘	6礫石灘	7開闊潮間帶	8遮蔽岩岸	9遮蔽潮間帶	10遮蔽濕地
0	0	1	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0

Here is the original data. We found that there are missing data. Therefore, we tried to fill it with ‘’reasonable value’’. By observing the original data, we can tell the regularity on those missing value. It always the season4 of the same location and county.

K Nearest Neighbors is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbors are classified. Based on the k nearest neighbor algorithm, we filled it according to the other 3 seasons.

Normalization. Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing. While we are training a neural network-based model, feature scaling helps us speed up the convergence rate.

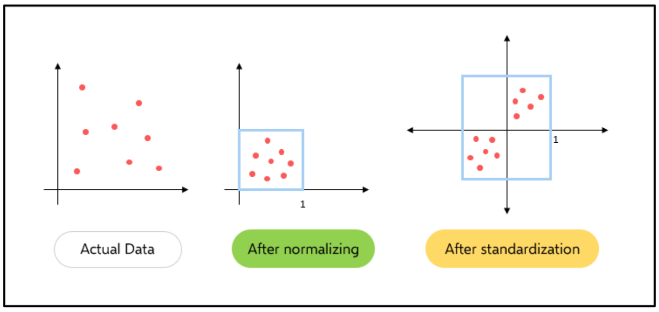


figure 2 The schematic diagram of data normalization

Feature selection. Features selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. When it comes to classification problem, entropy calculation is important. For example, while we are doing decision tree of student obesity and tried to choose the best matched feature. We found that student_ID might got the entropy which seems nice but it’s actually useless. Therefore, we drop out two kinds of feature County and location which is very similar to student_ID in the previous example.

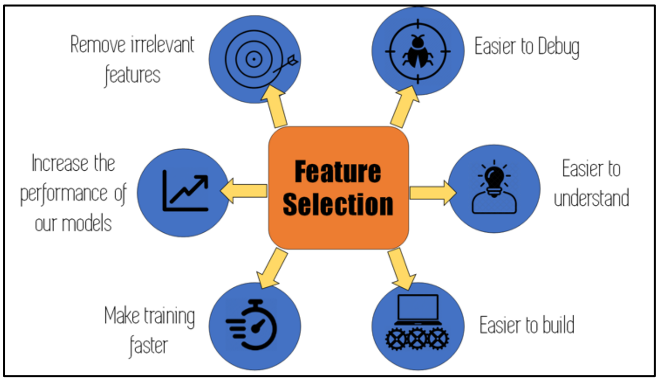


figure 3 The pros of feature selection

B. *Model and Input Features*

After data preprocessing, we construct a neural network (NN) model to predict the target. The NN model is quite simple, it is considered that the feature extraction and data preprocessing were more significant to model performance. Our model contains three fully-connected layer, each one has 256 neurons, using ReLU as activation function, finally coming with a softmax layer to predict the result.

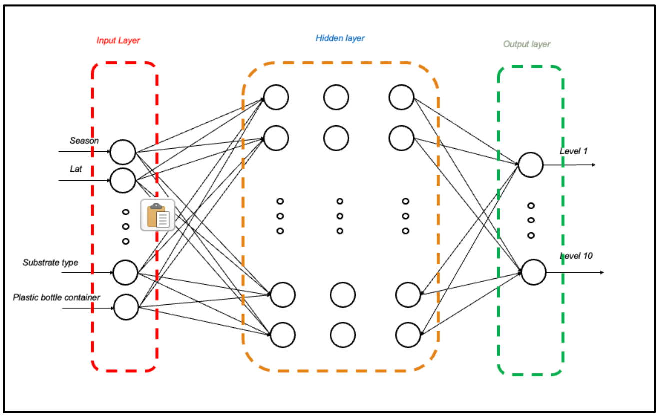


figure 4 The proposed neuron network

C. *Evaluation*

Evaluate the performance of modeling approach by Cohen's kappa, which is defined as:

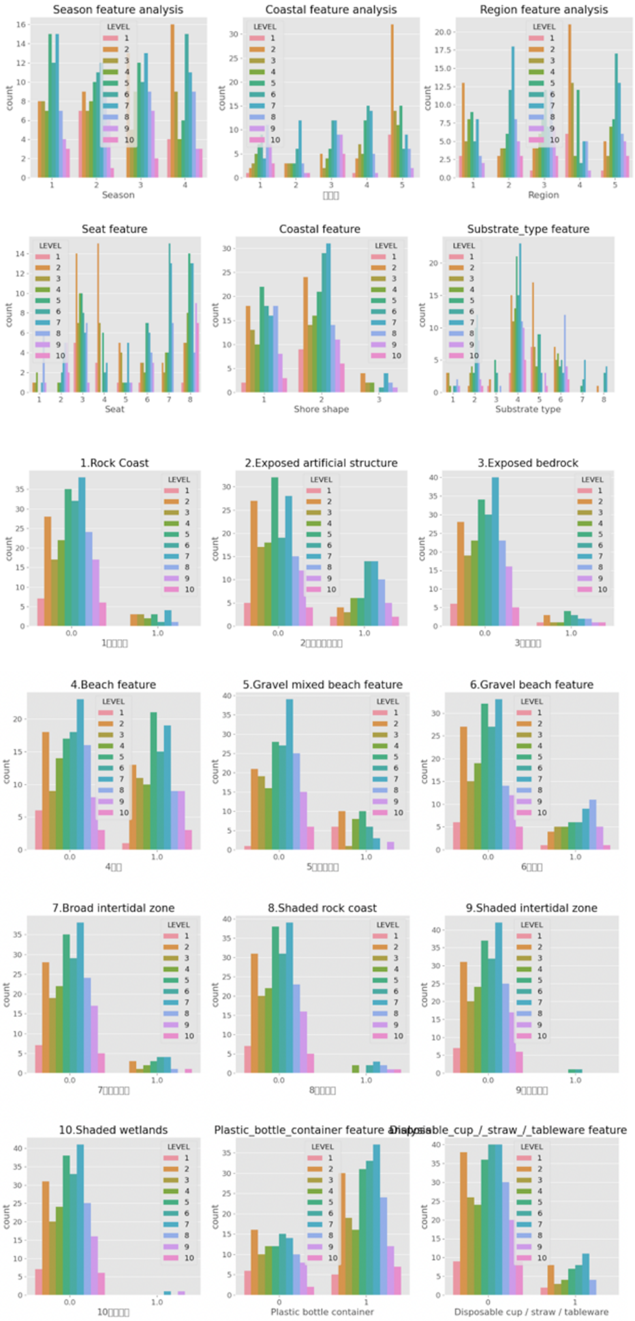
$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}$$

where $k = number\ of\ codes$ and w_{ij}, x_{ij}, m_{ij} are elements in the weight, observed, and expected matrices, respectively.

IV. Experiment

A. Feature Analysis

To improve the performance of the model, it's desirable to first analysis the input features. Here are the visualize analysis of all the input features.



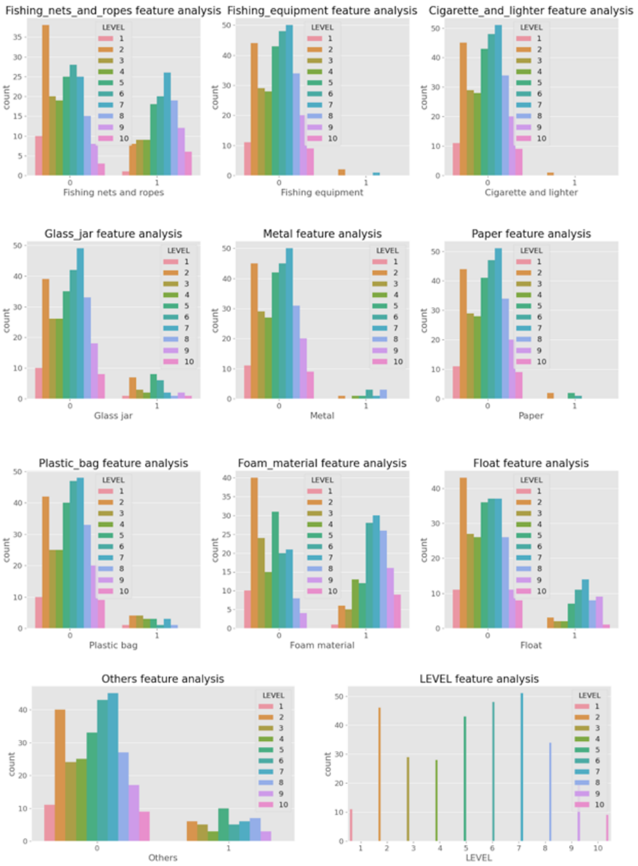


figure 5 Features analysis

B. Comparison

Different Method. We compared the best conhen's kappa scores for testing set of our different model in figure 6.

	Gradient Boosting	Random forest	Adaboost	Regression	Neural Network
Kapp	0.5839	0.5993	0.6184	0.4868	0.6713

figure 6 Comparing testing score for different types of model.

The neuron network model performs the best in our experiment.

With and Without Data Preprocessing. We compared the best conhen's kappa scores for testing set of our ANN model with and without data preprocessing in figure 7.

	Neural Network	Neural Network w/ Data prep
Kappa	0.6713	0.6814

figure 7 Comparing testing score for the model with and without data preprocessing.

With the data preprocessing the testing score can increase for about 0.01.

With and Without Early Stopping. We compared the best conhen's kappa scores for testing set of our ANN model with and without early stopping in figure.

	Neural Network	Neural Network w/ Early Stop
Kappa	0.6713	0.7085

figure 8 Comparing testing score for the model with and without Early stopping.

With the data preprocessing the testing score can increase for about 0.04.

C. Our Best model

Because our traditional NN model perform the best, we introduce the details of the model in the following article. It contains three fully-connected layer, each one has 256 neurons, using ReLU as activation function, finally coming with a softmax layer to predict the result.

Training Process. The accuracy and the loss of training and validation set is shown in the figure 9.

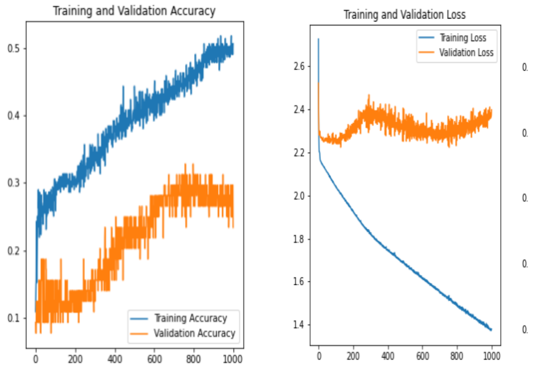


figure 9 The accuracy and the loss of training and validation

As we can see in the picture, the validation accuracy going down and the loss going up after 800 iterations, so we can easily conclude that the model is overfitting. To deal with the overfitting problem, we try to add the early stopping for regulization and the result is in figure 8.

Kappa Matrix. We also calculate the Kappa matrix between the predicted level (vertical axis) and the ground-truth level (horizontal axis), which can visualize the relationship between levels and the classification condition.

Kappa	0	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0	0
1	0	4	3	0	2	2	0	0	0	0	0
2	0	3	33	6	0	3	1	0	0	0	0
3	0	0	7	16	2	2	1	1	0	0	0
4	0	1	3	4	8	2	4	4	2	0	0
5	0	0	3	4	2	24	3	5	0	2	0
6	0	0	1	2	0	5	25	7	7	0	1
7	0	0	0	1	3	3	2	38	3	1	0
8	0	0	0	1	0	0	2	7	23	0	1
9	0	0	0	0	0	1	0	8	3	8	0
10	0	0	0	0	0	0	1	3	0	0	5

figure 10 Kappa matrix

As you can see in the figure 10, we expected the result to be all on diagonal area. The prediction of level 2,3,5,6,7,8 is better, and the prediction of 1, 4, 9, 10 is weak. Accoding to our feature analysis in figure 5, we concludes that 1, 4, 9, 10 perform weakly because the data are not balanced, so

V. Conclusion

In this project, we have proposed 5 types of models, which had its own pros as well as cons. In our settings, the regression had the worst prediction, on the contrary, the traditional artificial neuron network, had the best prediction on target. The prediction even improves after we adopt some common methods of data preprocessing such as filling up the missing value, data normalization, and features selecting. However, there still existed serious overfitting problem as we can see in the figure 9. To deal with the overfitting problem, we have tried to use early stopping to halt the training of neural network at the right time. By adopting the early stopping, the prediction on target had improved, which is shown in figure 8. Thus we can conclude that (1) When data is limited and the model are relatively shallow, data cleaning and preprocessing might be an essential step which will affected the result a lot. (2) Traditional Artificial Neural Network (ANN) still perform great on classification problem with efficient features and parameters.

VI. References

[1] Mijwel, Maad M. "Artificial neural networks advantages and disadvantages." Retrieved from LinkedIn https//www. linkedin. com/pulse/artificial-neuralnet Work (2018).

[2] Nazarenko, E., V. Varkentin, and T. Polyakova. "Features of Application of Machine Learning Methods for Classification of Network Traffic (Features, Advantages, Disadvantages)." *2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*. IEEE, 2019.