Modelling Chlorophyll-a Concentration using Deep Neural Networks considering Extreme Data Imbalance and Skewness

Jang-Ho Choi, Jiyong Kim, Jongho Won, Okgee Min

ETRI (Electronics and Telecommunications Research Institute), Daejeon, Republic of Korea

{janghochoi, kjy, jhwon, ogmin}@etri.re.kr

Abstract— **Algal bloom has been a serious problem, as some of algae such as cyanobacteria produce toxic wastes. Chlorophyll-a has been one of the primary indicator of algal bloom; however, it is difficult to model to forecast due to scarceness of the events. Since canonical machine learning algorithms assume balanced datasets, data imbalance of the Chlorophyll-a concentration must be visited for accurate prediction. In this paper, we present a convolutional neural network model to predict Chlorophyll-a concentration, handling its data imbalance and skewness. The experiment results show that proper data transformation and oversampling can improve prediction accuracy, especially in rare-event regions.**

Keywords— **Regression, Neural Network, Sensor-data regression, Data Imbalance, Data Skewness, Algal Bloom, Water Quality**

I. Introduction

For last few decades, algal bloom has been a serious problem, leading imbalance of organisms in the water. Severe algal bloom directly affects water quality as some of algae such as cyanobacteria produce toxic that are harmful to other species—including humans. Hence, the Korean Government has been operating the algal warning system since 1998 based on chlorophyll-a concentration and the number of cyanophyceae cells. Although using the concentration of chlorophyll-a as the primary indicator has been debatable, it is still widely used as it can be measured at low cost, whereas counting cyanophyceae cells incur considerable human labor.

There are two main approaches in forecasting algal blooms: simulation-based and machine learning-based approaches. Simulation-based approaches—though, may be accurate—must be carefully designed regionally by domain experts often with extensive information. Machine learning-based approaches, on the other hand, are relatively cheaper and easier to compute, while generating reasonable—often better—performance. The most advanced method in machine learning-based approaches to model chlorophyll-a concentration is using deep neural networks, inferring complex non-linear relationships to the environment variables.

Although deep neural network is capable to reflect non-linear relationships between input and target data, it requires a considerable amount of data from various circumstances to represent well. Unfortunately, water quality datasets are not only scarce, but also skewed and imbalanced due to their nature. For instance, algal bloom occurs extensively only during the summer, while these minority events are the important phenomena that must be forecasted. In fact, the data imbalance is significantly extreme that data, in which the concentration of Chlorophyll-a is above 100 mg/m^3 (categorized as algal bloom) take only 0.14% of the whole data, whereas 87.68% of them are below 15 mg/m^3 (categorized as safe level).

Hence, in this paper, we propose a deep neural network model to forecast chlorophyll-a concentration, considering its data imbalance and skewness. To mitigate extreme data skewness and imbalance, we apply log transformation and oversampling techniques; then, we conduct several experiments to analyze the effects of the techniques. Lastly, we present model performance in two metrics, the coefficient of determination (R^2) and mean squared error (MSE) for comparisons.

II. Handling Data Imbalance

In this paper, we utilize two distinct techniques to handle data imbalance. Since the data distribution of Chlorophyll-a concentration is extremely skewed, we applied log-transformation to make distribution conform to normality. Log transformation is a simple but useful technique to increase validity of the applied analysis, as many of the methods assume data normality. The example of log transformation is shown in Figure 3.

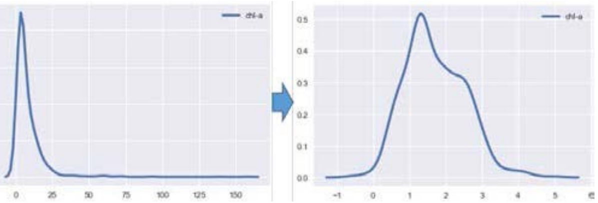

Figure 1. Appling Log Transformation

Another technique we applied to handle data imbalance is oversampling. One of the most intuitive data-level techniques is to either undersample the majority class or oversample the majority class [1]. We should avoid undersampling because we may lose useful information especially when the size of dataset is very limited [2]. On the other hand, oversampling can preserve all the information currently available. However, oversampling increases computation in training, though the burden is quite small as the original dataset itself is not large.

The simplest oversampling method would be replicating or randomly sampling minority instances; however, it may lead to overfitting [3] and potentially create confusion in training [4][5]. Hence, in this work, we adopted Synthetic Minority Oversampling TEchnique (SMOTE) [6], in order to increase samples in minority regions. The main idea of SMOTE is to generate new instances using Euclidean distance vectors to the nearest neighbors. It is originally proposed for classification problem so that we need to adjust experiment setting in order to apply SMOTE.

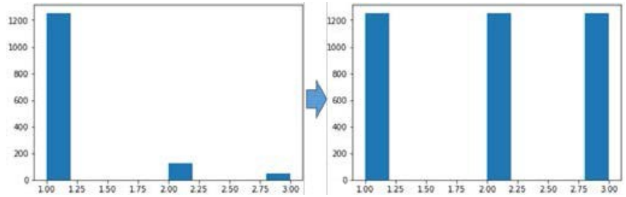

Figure 2. Oversamping minority classes to make balanced dataset using SMOTE

Since the data distribution of Chlorophyll-a concentration is skewed and the size of dataset is small, it is quite challenging to convert to a set of classes. If we use small-sized bins—such as 5 to 10mg/m^3—most of the bins would be empty and their ranges will not be reflected in sampling and/or learning. Therefore, we should use a few, larger bins instead, decided to match them with the stages of the algal warning system—that is normal, caution, warning, and algal bloom alert.

Nevertheless, the goal of our prediction model is to forecast Chlorophyll-a concentration, not their stages in terms of classes. True, one may predict the concentration by taking mean value of the bean, if the bin width is small enough. Due to the large bin width, however, it would often generate significant errors. Instead, we concatenate input features with target variable so that both of the variables are generated in the oversampling process. Then, we split the input features and target variable for training the regression model. Without question, oversampling techniques should only be applied to training set. Therefore, in case of cross-validation, the sampling process must reside in the iterations, avoiding overfitting and misleading results.

III. Experiment

In this section, we demonstrate the effects of data imbalance in regression problem. We have conducted four experiments to compare the effects of mitigation techniques on skewed data.

A. Dataset and Preprocessing

For the experiments, we used daily data collected from water quality monitoring station at Daechungho lake during 2012 to 2018 that is opened to the public by Korea's National Institute of Environmental Research (NIER). The collected data consist of eight measurements: water temperature, pH, electrical conductivity, dissolved oxygen, total organic carbons, total nitrogen, total phosphorus and chlorophyll-a. The data distribution of target variable, chlorophyll-a is extremely skewed that the concentration of chlorophyll-a above 100mg/m^3 takes only 0.14% of the whole dataset. The histogram of Chlorophyll-a Concentration is shown in Figure 1.

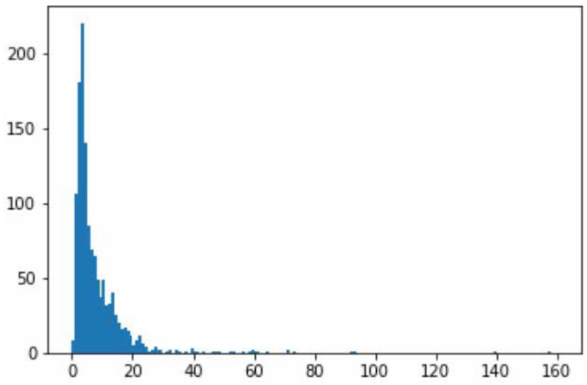

Figure 3. Histogram of Chlorophyll-a Concentration (mg/m^3, Binwidth=1)

There are a large number of missing data in the dataset. We removed instances whose number of missing features are more than four. We also dropped instances if at least one of their features is missing for 5 or more days in a row; then, we linearly interpolate the rest of the missing features to make a complete dataset. To fully utilize the given dataset, we create (N-1)-day overlapping time-series data, generating 1295 instances of N x M features.

B. Model Architecture and Experiment Setting

The goal is to accurately predict the concentration of chlorophyll-a in 7 days. The authors in [7] demonstrate that a convolutional neural network (CNN) model is a well-suited model for regression-type problems, capturing dependencies in and between time-series. Hence, we decided to build CNN-based model, whose architecture is portrayed in Figure 2.

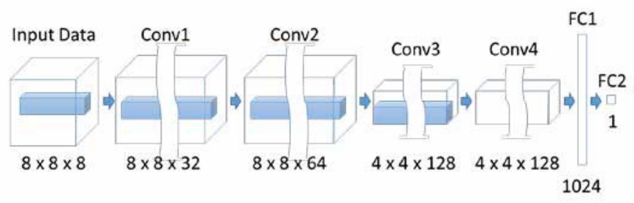

Figure 4. Convolutional Neural Network Model Architecture

The proposed model consists of four convolutional layers and two fully-connected layer including the output layer. Each convolutional layer has 32, 64, or 128 filters with 3x3 kernel, followed by ReLu activations. We also added a max-pooling layer with pooling size of 2x2 and 25% dropout every after two convolutional layers. The first fully-connected layer has

1024 units, also followed by ReLu activation and 50% dropout to the output layer. In total, the prediction model has about 76k parameters.

Since the target dataset is small and very skewed, we applied cross validation with 10-folds. Cross validation is rather mandatory for such dataset, as extreme regions can be eliminated from test set with high probability, leading the results invalid and unreliable. In fact, we observed a large variation of performance during the 10-fold cross-validation.

C. *Results and Comparison*

In this section, we demonstrate the effects of two techniques in modelling imbalanced data. We used 10-fold cross-validation to evaluate the performance with two metrics: the coefficient of determination (R^2) and mean squared error (MSE). The coefficient of determination is a common metric to validate regression model performance, depicting how close the data are to the fitted regression. MSE is also a good estimator that measures the average of the squares of the errors between unseen test data and predicted data. Since R^2 and MSE on training data is nearly 1 and 0 accordingly, we present scores on the test set only.

We have conducted four experiments with the following methods for comparison:

- baseline, without log transformation and oversampling
- log transformation
- oversampling using SMOTE (k=5)
- log transformation and oversampling using SMOTE

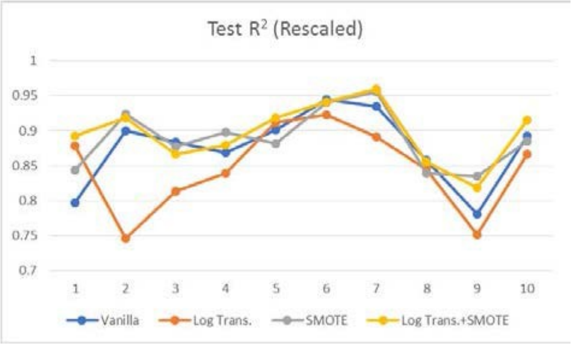

Figure 5. R^2 on Test Set (10-fold CV)

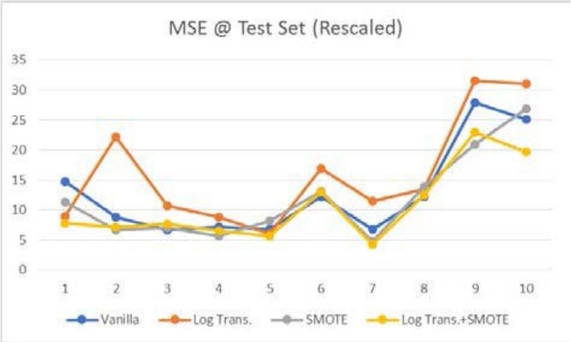

Figure 6. Mean Squared Error on Test Set (10-fold CV)

Figure 5 and 6 show the R^2 and MSE of the experiments accordingly. The result is quite interesting that vanilla mode without log transformation and oversampling outperforms one with log transformation. Log transformation provides the best performance—the highest R^2 and lowest MSE—if measured before rescaling the test data. In other words, the transformed error is small, fitting regression well; however, the error also propagates as the predicted value is rescaled.

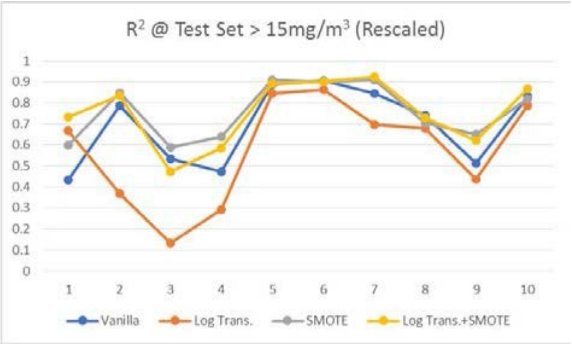

Figure 7. R^2 on Test Set > 15mg/m^3 (10-fold CV)

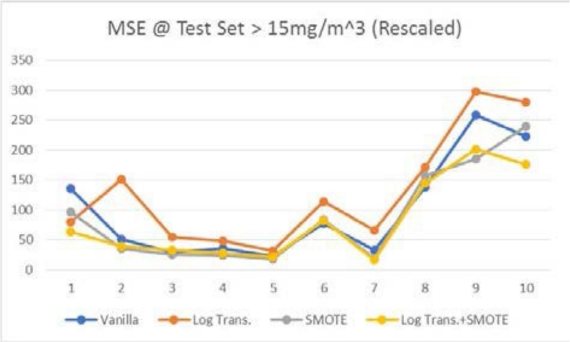

Figure 8. Mean Squared Error on Test Set > 15mg/m^3 (10-fold CV)

The experiment result shows that the oversampling technique, SMOTE works surprisingly well even on the extremely imbalanced dataset, despite the fact that more than half of the training dataset were synthetically generated. Oversampling mode outperformed vanilla mode both in R^2 and MSE in most iterations. The result can be interpreted as that the generated synthetic data help the model avoid being biased toward the majority regions.

TABLE 1. PERFORMANCE COMPARISON ON WHOLE TEST SET

	Rescaled R Squared			
	Vanilla	**Log Trans.**	**SMOTE**	**LT + SMOTE**
average	0.87599979	0.84646486	0.88787952	**0.89645384**
min	0.7807487	0.74654529	**0.83493511**	0.81885176
max	0.94449447	0.92304627	0.95531663	**0.95938027**
	Rescaled MSE			
	Vanilla	**Log Trans.**	**SMOTE**	**LT + SMOTE**
average	12.850588	16.1193393	11.8379616	**10.7345853**
min	6.7118805	6.04734636	4.70396465	**4.27617197**
max	27.8473856	31.6011565	26.8849693	**23.0078686**

Another surprising result we discovered is that applying oversampling technique together with log transformation generate better—or at least comparable—performance. Our interpretation is that log transformation alters the skewed distribution of synthetic data in each bins to conform towards normality, leading better performance. We will conduct further experiments to validate our hypothesis in the future.

TABLE 2. PERFORMANCE COMPARISON ON TEST SET >= 25MG/M^3

	Rescaled R Squared			
	Vanilla	**Log Trans.**	**SMOTE**	**LT + SMOTE**
average	0.6958938	0.57720683	**0.7572996**	**0.75571792**
min	0.43423147	0.13239613	**0.58755125**	0.47152785
max	0.90787247	0.86456359	0.91167817	**0.92351488**
	Rescaled MSE			
	Vanilla	**Log Trans.**	**SMOTE**	**LT + SMOTE**
average	100.489066	129.524892	88.7204496	**80.9447425**
min	22.6287369	31.8372526	18.1415467	**16.6756251**
max	258.648759	298.052054	239.651678	**201.59149**

The performance gap diverges when we limit the test set range from 15mg/m^3, which is the beginning of the minority regions. Figure 7 and 8 show the R^2 and MSE of test set in the regions, accordingly. The average, minimum, and maximum statistics of the experiments are also illustrated in Table 1 and 2. In general, the combination of log transformation and SMOTE generates the best performance.

IV. CONCLUSIONS

We present a convolutional neural network model to predict Chlorophyll-a concentration, handling its data imbalance and skewness. To validate our approach, we demonstrate the effects of data transformation and oversampling techniques on extremely skewed data. In addition, we describe how to apply oversampling techniques in regression problem—while avoiding overfitting—when data distribution of the target variable is largely imbalanced. The experiments show that log transformation and oversampling technique together help improve performance of the prediction model, especially in minority regions.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00219, Space-time complex artificial intelligence blue-green algae prediction technology based on direct-readable water quality complex sensor and hyperspectral image)

REFERENCES

[1] Haibo He and Edwardo A. Garcia. (2009). Learning from imbalanced data. *IEEE Transactions On Knowledge And Data Engineering*, 21(9):1263–1284.

[2] Bartosz Krawczyk, Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence*, November 2016, Volume 5, Issue 4

[3] Salman H. Khan et al., Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, Issue: 8 (2015)

[4] N. Japkowicz, "Class Imbalances: Are We Focusing on the Right Issue?" *Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II*, 2003.

[5] C. Drummond and R.C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under Sampling Beats Over-Sampling," *Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II*, 2003

[6] N. V. Chawla et al., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Vol. 16 (2002)

[7] Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time-series forecasting with convolutional neural networks. *arXiv preprint* arXiv:1703.04691.

Jang-Ho Choi received his B.S. degree in Computer Science & Business Administration in 2010 from the University of Southern California, Los Angeles, USA. He received his MS in Computer Science from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2013. He joined SW & Contents Research Laboratory at ETRI as a researcher in 2013. His research interests include machine learning, deep learning, big data analysis, and artificial intelligence.

Kim Jiyong received the M.S. degree from Seoul National University, Korea, in 1997. He is a principal researcher at Electronics and Telecommunications Research Institute(ETRI). His research topics include big data analysis, machine learning and IoT. Currently he is in charge of the project, "Space-time complex analysis technology for blue-green algae prediction".

Won Jongho received the M.S degree in computer science from Inha University, Korea, in 1989, and the Ph.D degree in computer engineering from Univ. of Texas at Arlingthon, U.S.A., in 1997. In 1998, he joined AT&T Lab in Middletown, NJ, U.S.A. as a senior research staff. Since 2000, he has been with Electronics and Telecommunications Research Institute(ETRI), Korea, as a principal researcher. His research topics include big data analysis, machine learning and data management.

Ok-gee Min received her BS and MS degrees in Computer Science and Statistics from Chungnam National University, Daejeon, Rep. of Korea in 1988 and 1992, respectively. She received her PhD in Computer Engineering from Chungnam National University in 2010. She is currently a director and principal researcher at ETRI, Daejeon, Rep. of Korea. Her research interests include data stream processing, big-data analysis, data intelligence, and cognitive architectures.