Machine Learning Approaches for Cyanobacteria Bloom Prediction using Metagenomic Sequence Data: a Case Study

JianDong Huang Huiru (Jane) Zheng* Haiying Wang

School of Computing
Ulster University
Jordanstown, County Antrim, Northern Ireland, UK
jd.huang@ulster.ac.uk, h.zheng@ulster.ac.uk*, hy.wang@ulster.ac.uk

Abstract Cyanobacteria bloom is a serious public health threat and a global challenge. Literature on the bloom prediction and forecasting has been accumulating and the emphasis appears to have been on the relation between the blooms and environmental factors, whilst the complexity of the bloom mechanism makes it difficult to reach adequate output of the models. Rapid development of next generation sequencing techniques provides a way in which comprehensive and quick examination of the microbial community can be achieved, especially for the bloom community structure. This facilitates using of merely the sequence data along with the machine learning techniques to predict and forecast the bloom occurrence. But there has been rare report on this theme in the literature. In this case study, machine learning approaches were applied with the metagenomic data as the only input (rather than with environmental data) to predict the cyanobacteria blooms. k-NN classification, SVM classification and k-means clustering were applied and their efficiencies were evaluated using relevant indices. Feature selection was performed and the yielded sub datasets were worked on seriatim. In the predicting experiment with k-NN approach, the final year's data among the 8 years OTU time series were used as target data and various combination of the preceding years' data were used as predictor data; the output came with the best values of 1.00 and 100% for the evaluation indices F1 score and sensitivity, specificity, precision, and accuracy, for the 7 preceding years' predictor input, among the experiment results. This case study demonstrated the feasibility of using machine learning approaches in the Cyanobacteria bloom prediction with only metagenomic sequence data, and the importance of feature selection processing in obtaining better output of the machine learning approaches. The metagenomic data based machine learning approaches are efficient, economic, and faster, possessing the advantage and potential for being adopted as a promising means in the bloom prediction practice.

Keywords Machine learning, Cyanobacteria blooms; OTU (Operational Taxonomic Unit)

This study was part of the "Wuhan lakes" project funded by Ulster University Global Challenges Research Pump-Priming Fund.

I. Introduction

Over last decades, a world–wide increase in the incidence of harmful cyanobacteria blooms has prompted a large amount of studies into the hazard, cause, and prediction of this phenomenon [1]-[4]. One category of prediction effort since early days has been laying emphasis on the influence of environmental factors as physical, chemical and biological parameters on the forming and trigging of the blooms. For example, Yabunaka *et al.* used nutrients and physic-chemical conditions (such as nitrogen, phosphorous, water temperature and transparency, dissolved oxygen, pH and so on) as the input of their Artificial Neural Network models for prediction of the blooms in Tolo Harbour, Hong Kong [5]. In a study by Wu *et al.*, environmental factors as input variables into an EFDC (Environmental Fluid Dynamics Code) model for chlorophyll-a simulation and algal bloom prediction included nutrition parameters and water physical & chemical parameters, in the Daoxiang Lake, Beijing [6]. They reported that the average algal bloom prediction accuracy was 63.43%. Vilán *et al.* built a cyanotoxin diagnostic model by using machine learning techniques in the Trasona reservoir in Northern Spain, the input variables were a number of biological and physical-chemical variables, and the former included the microcystis and other cyanobacteria species [7]. Li *et al.* applied a coupled hydrodynamic–algal biomass model for forecasting short-term cyanobacterial blooms in Lake Taihu; the model was applied to predict the occurrences of the algae blooms in Lake Taihu during April to September in 2009 and 2010. The observations of chlorophyll a concentrations were used to calibrate the model [8]. They stated that independent evaluations from remote sensing images and boat survey data showed that the accuracy of the bloom forecasts was more than 80%. A recent study by Lou *at el*. selected 15 variables such as alkalinity, bicarbonate (HCO_3^-), dissolved oxygen (DO), total nitrogen (TN), turbidity, conductivity, nitrate, suspended solid (SS), and total organic carbon (TOC) for their hybrid intelligent model simulation and prediction of the blooms [9]. The prediction and forecast powers were estimated at approximately 0.767 and 0.876 respectively. In general, these category of environmental factor driven modelling methods involve deploying large number of varies type of instruments for parameter collection and analysis;

at the same time, more adequate prediction performance has been expected as well.

With the rapid development of next generation sequencing techniques, metagenomic data analysis has been applied to the Cyanobacteria bloom research. Based on the system of 16S rRNA gene, new generation high-throughput sequencing techniques facilitates examination of the composition of the microbial community comprehensively and quickly in different habitats, enabling insight into profiles of the community composition [10]-[12]. Application of metagenomics in investigating the genetic and metabolic diversity of the mixed populations helps understand the interactions of different microbial populations and their functions in the blooming process. Nevertheless, the reality of quick detection of the OTU feature of the microbial community and the possibility of using merely the time course sequence data for the bloom prediction underpinned by machine learning techniques, prompt the notion of machine learning solution for a higher accuracy performance of prediction with a lower cost in terms of only requesting the sequencing data yielded from collected water samples instead of acquiring lots of environmental parameters. A relevant advance has been made in the recent study by Tromas *et al.*, where they predicted cyanobacterial blooms in an 8-tear amplicon sequencing time course [13]. In the study, they predicted the start date of a bloom with 78-92% accuracy, and concluded that sequence data was a better predictor than environmental variables. Incited by the notion and the advance, our work described here examined the performance of three machine learning approaches covering both supervised and unsupervised methods, k-nearest neighbours (k-NN), Support Vector Machine (SVM), and k-means clustering approach, to demonstrate the feasibility of machine learning approaches in the cyanobacteria bloom prediction practice.

II. Dataset and Methods

A. Temporal series dataset

The dataset used in this study was the output of a deep 16S amplicon sequencing analysis for samples collected from the photic zone (0-1 metre depth), Lake Champlain, Quebec, Canada, from 2006 to 2013 between April and November of each year, related to the study by Nicolas et al [13]. The sampling spanned multiple bloom events. Samples were acquired from both littoral and pelagic zones. The sequence analysis and OTU picking yielded a final data set of 135 samples. The data set was clustered into 4061 OTUs. The bloom and non-bloom samples were labeled (http://www.nature.com/ismej) and this feature were used in the further analysis. In the 135 samples, there were 33 "bloom" samples and 102 "non-bloom" samples.

B. Dataset normalization

As a common practice, the OTU dataset was processed with normalization firstly, to convert all the values into [0, 1] interval for later calculation/modelling. The formula is:

$$Z_{ij} = (X_{ij} - X_{j\_min}) / (X_{j\_max} - X_{j\_min}) \qquad (1)$$

where, Z_{ij} is normalized value, X_{ij} is the value in the original dataset, X_{j_max} and X_{j_min} are maximum and minimum of the j-th variable (feature) respectively.

C. Feature selection

Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction in machine learning and statistics. The reasons for feature selection are four folds: to simplify models to make them easier to interpret by researchers/users [14], to shorter training times, to avoid the curse of dimensionality, and for enhanced generalization by reducing overfitting [15]. The central premise when using a feature selection technique is that the data contains many features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information [15]. In this study, feature selection was performed on the OTU data set with Relief algorithm. A feature subset selection is a task of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept [16]-[18]. Relief algorithms are general and successful attribute estimators and are especially good in detecting conditional dependencies [18]. It was realized in Matlab R2017a (9.2.0.556.344) in the processing. Eleven subsets of the data were obtained according to the output ranks of the feature selection, for further analysis.

D. k-nearest neighbours (k-NN), Support Vector Machine (SVM), and k-means Clustering approaches

k-nearest neighbours algorithm (k-NN) is a non-parametric method used for classification and regression. For the classification, k-NN classifier is to classify unlabelled observations by assigning them to the class of the most similar labeled examples. The input consists of the k closest training examples in the feature space [19]-[20]. In our work, *fitcknn* with the function of optimizing fitted k-NN classifier was deployed, which optimize hyperparameters automatically for the k-NN analysing in Matlab 2017a. Significantly, the value k was automatically determined by the optimization process.

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane [21-22]. Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall (Wikipedia, the free encyclopedia). This study required to separate the dataset points into two classes for bloom and non-bloom examining only, and this was realized in Matlab package.

k-means clustering analysis groups a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups (clusters) [23]. k-means clustering aims to partition n observations into k clusters

in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. For grouping the bloom and non-bloom samples, k-mean clustering was carried out in Matlab with k=2. For each trial run, the output were the coordinates of two centroid points for the two clusters; one of which was determined as belonging to the blooming cluster by its smaller sum or average of the distance to the bloom-labeled sample points. In blooming prediction practice, the centroid point of the blooming cluster can be fixed by analysing one (or a few) sample(s) and determining its bloom or non-bloom belonging (whether or not containing cyanobacteria phylum), then the clusters will be identified.

In the machine learning approach, 70% of the dataset samples (N=95) were used for training and establishing the models, and the rest 30% (N=40) were used for testing. The training sets were selected from the front 70% of the temporal series, for the purpose of a direct examining against the real blooming status label, of the performance of the models, especially their capacity in potential practical usage.

For a further investigation of the efficiency of the machine learning approaches for forecasting (hindcating in this case study) of the cyanobacteria blooms, we also took the final year's bloom and non-bloom data in the dataset (data of the year 2013) as target variables, and a series of combination of the preceding years' data as predictor variables, to perform k-NN analysis and compared the output of the hindcasting performance. There were 7 dataset for input of the experiments: data for the year 2012 (1 year), for the year 2012 and 2011 (2 years), for the year 2012, 2011, and 2010 (3 years)… and then, for the year 2012 back successively to 2006 (7 years).

E. Modelling result evaluation indices

Terms and indices were applied in the model evaluation as below in Table 1, where PP is the total number of samples labeled "bloom" and NN is the total number of samples labeled "non-bloom", for a selected dataset. For the dataset containing the whole of 135 samples, PP=33 and NN=102, but for those sub-sets, PP and NN values varies. Noticing that the bloom and non-bloom samples are quite unbalanced in their numbers, an index F1 score, which is a measure of a test's accuracy, was specially selected to describe the performance of the different methods. F1 score considers both the precision PPV and the sensitivity TPR of the test to compute the score, here PPV is the number of correct positive results divided by the number of all positive results and TPR is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score is the harmonic average of the precision and sensitivity, an F1 score reaches its best value at 1 (perfect precision and sensitivity or recall), and worst at 0.

Table 1. Terms and indices applied in the model evaluation

Term or index	Meaning or formula
TP	True Positive (hit)
TN	True Negative (Correct rejection)
FP	False Positive (false alarm)
TN	True Negative (Correct rejection)
Sensitivity (True Positive Rate, TPR)	TPR = TP / PP (PP: sum of condition positive)
Specificity (SPC, True Negative Rate, TNR)	TNR = TN / NN (NN: sum of condition negative)
Precision (Positive Predictive Value, PPV)	PPV = TP / (TP + FP)
F1 score (the harmonic mean of precision and sensitivity)	F1 =2*TPR*PPV / (TPR+PPV)
Accuracy (ACC)	ACC = (TP + TN) / (PP + NN)
Silhouette coefficient **(for evaluation of output of k-means clustering)**	$S_i = (b_i - a_i) / \max(a_i, b_i)$ a_i: the average of the distance the ith point to other points in its own group **b_i: the average of the distance the ith point to the points in the opposite group**

which were determined by the *Relieff* function in the Matlab package. Eleven sub datasets were formed through removing some features (variables) in the original dataset, with their weights (and associated ranks) lower than the given level. For example, for a weight level 0.01, there were 503 features (variables) whose weights found to be above 0.01, and then these features were selected to form the subset sub0503. Table 2 gives details of all the sub datasets. Among the 11 subset, the sub50 and sub37 yielded poor results in the analysis thus they were not used in the later work. After the feature selection, number of the variables used for modelling was sharply reduced and this would avail the further processing.

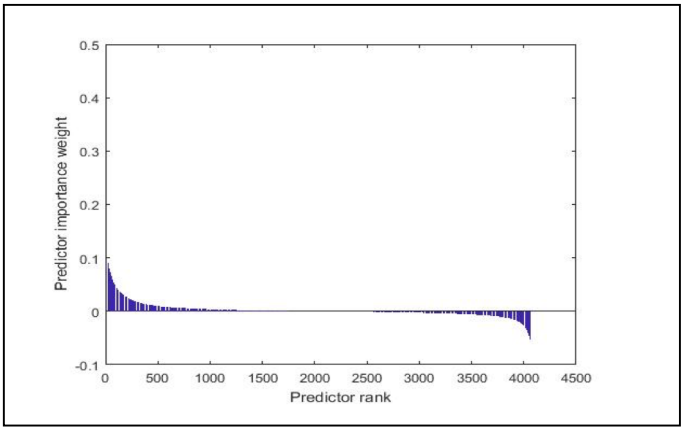

Figure 1. Feature selection for the dataset

III. Results and discussion

Figure 1 shows the feature selection result. Depending on the feature selection output Predictor Importance Weight,

To investigate the performance and efficiency of k-NN, SVM classification and k-means clustering in the prediction, all of the sub datasets were fed and undergone the training and testing in turn. For potential practical application of the established models, the training set were selected from the up-front 70% of each sub dataset and the rest 30% were used for testing, or, more significantly, hindcasting.

Table 2. Selection of features according to the output of Feature Selection

Weight (between -1 and +1) level	Number of features (variables) with weight greater than the weight level in the full dataset	Number of features(variables) selected in each subset	Percentage of the total variables (%)
0.00	1968	Subset 1: 1968	48.4610
0.01	503	Subset 2: 503	12.3861
0.02	272	Subset 3: 272	6.6979
0.03	181	Subset 4: 181	4.4570
0.04	119	Subset 5: 119	2.9303
0.05	89	Subset 6: 89	2.1916
0.06	65	Subset 7: 65	1.6006
0.07	50	Subset 8: 50	1.2312
0.08	37	Subset 9: 37	0.9111
0.09	27	Subset 10: 27	0.6649
0.10	18	Subset 11: 18	0.4432

Table 3, Table 4 and Table 5 show the results of the classification and the clustering processing respectively.

For the k-means analysis, a Silhouette plot was created for visualized examination of the performance, shown in Figure 2. The Silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. Its value ranges in [-1, +1], and a high Silhouette value means the point is well matched to its own cluster. Figure 2 shows that there was no

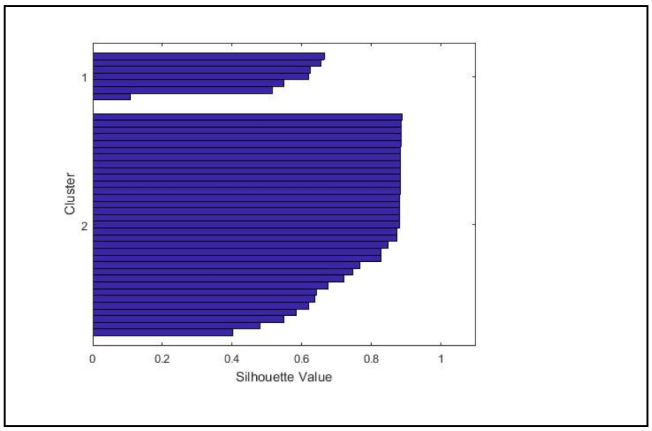

Figure 2. Silhouette plot of the k-means clustering in this study

negative values appeared for any point in the clustering, indicating that the grouping solution was adequate.

Table 3 shows the performance of the k-means clustering approach for 9 data subset, measured by 5 indices. Subset sub1968 had best F1 score and the values for the others are all above 0.70. As an unsupervised approach, k-means clustering (k=2) in the bloom study does not need the bloom and non-bloom labels in its model establishing stage. After having determined the two centriod points in the multi-dimensional space, the prediction will be to decide a new sample's belonging between the two clusters by calculating the distance of the point (the sample) to the two centroid points and grouping it to the one with shorter distance. In practice, it would be handy to determine which cluster is the "bloom cluster" by analysing one or a few samples to find out the cyanobacteria-containing information then assign the label. This distinctive character then appears to make the k-means clustering approach a convenient and practical tool for the bloom monitoring and predicting with proper quality, although the performance of this method has appeared not as good as the other supervised methods used in this case study (Tables 4, 5 and 6).

In Table 4, the measures for the SVM performance show that the majority of the subsets have their F1 scores larger than 0.75, and the subset0027 and sub0018 have the highest F1 scores (0.8696 and 0.8462). Other indices also indicate that the SVM approach, originally designed for binary classification, appears to be a suitable means for the bloom prediction as well.

In Table 5 it can be seen that k-NN approach (the value of k was automatically determined in the *fitcknn* function processing in Matlab) is superior to the other machine learning methods in this case study, demonstrated by the indices. All the F1 scores are above 0.80 and the best ones are 0.9231 for the subset sub0089 and sub0181. The other parameters are superb, and it is noticeable that the processing with the total dataset (all 4061 variables inclusive) did not yield higher values of the indices but relatively lowest values.

The average index values of the subsets for the three machine learning methods are plotted in Figure 3. The k-means approach corresponds to good values of precision and specificity, but the low sensitivity value leads to its low F1 score; the k-NN approach has the highest F1 score and in general superior to the other methods.

To examine whether there were significant difference among the output of the three machine learning models in terms of the F1 measures, a statistical ANOVA (Analysis Of Variance) test was performed in Matlab with the *anova1* function. The result showed an F value of 14.79, and a P value (Probability > F) of 6.51862e-0.05. This exhibited that the difference were significant and the k-NN model was relatively the best one with its performance.

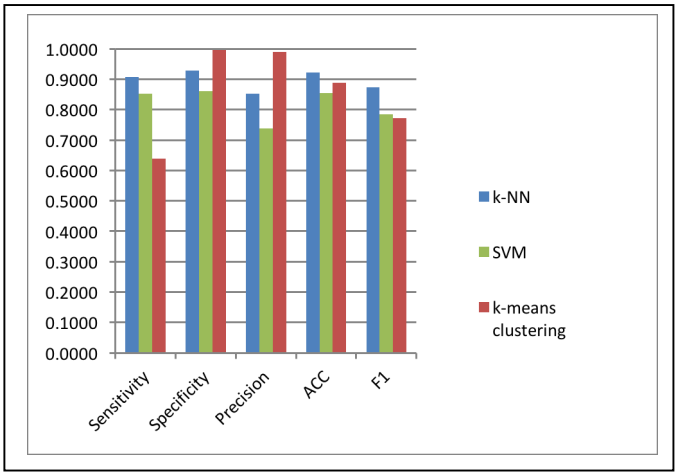

Figure 3. Average values of the five indices showing the performance of the three machine learning methods in the cyanobacteria bloom hindcasting.

Table 3. Performance of k-means clustering approach in hindcasting the blooms (see Table 1 for the definition of the terms herein)

Subset	TP	FP	TN	FN	Sensitivity (TPR, %)	Specificity (SPC, TNR, %)	Precision (PPV, %)	ACC (%)	F1
sub0018	7	0	28	5	58.33	100.00	100.00	87.50	0.7368
sub0027	7	0	28	5	58.33	100.00	100.00	87.50	0.7368
sub0065	7	0	28	5	58.33	100.00	100.00	87.50	0.7368
sub0089	7	0	28	5	58.33	100.00	100.00	87.50	0.7368
sub0119	7	0	28	5	58.33	100.00	100.00	87.50	0.7368
sub0181	8	0	28	4	66.67	100.00	100.00	90.00	**0.8000**
sub0272	8	0	28	4	66.67	100.00	100.00	90.00	0.8000
sub0503	8	0	28	4	66.67	100.00	100.00	90.00	0.8000
sub1968	10	1	27	2	**83.33**	96.43	90.91	**92.50**	**0.8696**

* Number of samples used for the hindcasting: N=40, N_{bloom} = 12, $N_{non\text{-}bloom}$ = 28.

Table 4. Performance of SVM approach in hindcasting the blooms (see Table 1 for the definition of the terms herein)

Subset	TP	FP	TN	FN	Sensitivity (TPR, %)	Specificity (SPC, TNR, %)	Precision (PPV, %)	ACC (%)	F1 score
sub0018	11	3	25	1	91.67	89.29	78.57	90.00	**0.8462**
sub0027	10	1	27	2	83.33	96.43	90.91	92.50	**0.8696**
sub0065	10	5	23	2	83.33	82.14	66.67	82.50	0.7407
sub0089	10	4	24	2	83.33	85.71	71.43	85.00	0.7692
sub0119	10	4	24	2	83.33	85.71	71.43	85.00	0.7692
sub0181	12	8	20	0	100.00	71.43	60.00	80.00	0.7500
sub0272	12	7	21	0	100.00	75.00	63.16	82.50	0.7692
sub0503	10	4	24	2	83.33	85.71	71.43	85.00	0.7692
sub1968	9	2	26	3	75.00	92.86	81.82	87.50	0.7826

* Number of samples used for the hindcasting: N=40, N_{bloom} = 12, $N_{non\text{-}bloom}$ = 28.

Table 5. Performance of k-NN approach in hindcasting the blooms (see Table 1 for the definition of the terms herein)

Subset	TP	FP	TN	FN	Sensitivity (TPR, %)	Specificity (SPC, TNR, %)	Precision (PPV, %)	ACC (%)	F1 score
sub0018	10	3	25	2	83.33	89.29	76.92	87.50	0.8000
sub0027	10	1	27	2	83.33	96.43	90.91	92.50	0.8696
sub0065	12	3	25	0	100.00	89.29	80.00	92.50	0.8889
sub0089	12	2	26	0	100.00	92.86	85.71	95.00	**0.9231**
sub0119	9	0	28	3	75.00	100.00	100.00	92.50	0.8571
sub0181	12	2	26	0	100.00	92.86	85.71	95.00	**0.9231**
sub0272	11	2	26	1	91.67	92.86	84.62	92.50	0.8800
sub0503	11	2	26	1	91.67	92.86	84.62	92.50	0.8800
sub1968	11	3	25	1	91.67	89.29	78.57	90.00	0.8462
Total4061	9	1	27	3	75.00	96.43	90.00	90.00	0.8182

* Number of samples used for the hindcasting: N=40, N_{bloom} = 12, $N_{non\text{-}bloom}$ = 28.

Given that the k-NN approach performed best in this case study, a further investigation with k-NN was attempted using one year's data (the final year 2013 in the whole dataset) as the testing or hindcasting dataset and the preceding year(s) data as the training data, to examine the efficiency of the method in bloom prediction. The input-output pairs were from 1 year (2012), 2 years (2012 and 2011), 3 years (2012, 2011, 2010), 4 years (2012, 2011, 2010, 2009), ……, to 7 years (2012, 2011, … 2006), against the year 2013 whose data was used as the testing dataset. Table 6 shows the results selected from the best performed output. Because the F1 scores were all lower than 0.80 for those datasets formed by sum of 3 or less than 3 years (for example, F1=0.25 were the output for some of the one year dataset), only those with F1 score larger than 0.80 were listed in the Table and the rest were not accepted for further analysis. The column "Number of years" indicates how many preceding years' data were used as the input dataset. Because it was the case that in many runs different years' dataset yielded the same indices values, then these sorts of output were placed in the same row of the table. For example, in the row for sub0018, the column "Number of years" shows 4, 6, 7, corresponds to the situation that the 4, 6, and 7 preceding years' dataset respectively were the input dataset, y. They yielded the same value for an index such as Sensitivity or F1 and so on.

In Table 6, subsets sub0119 and sub0181 correspond to the best value 100% and 1.0, of all the indices, with k-NN approach. This demonstrates that the k-NN can be a powerful tool for the bloom forecasting, noticing that the targeting dataset is "new" to the established model, and the high values of the indices, comparing especially with the reported performance of afore mentioned environmental factor data driven models.

The indices listed in Table 6 were selected from the output of the trials with best performance, and it can be seen that the sum of the preceding years range from for 4, 5, 6 and 7 years. But the most of the results were yielded from the input of 7-years datasets. This appears to be in accordance with the common knowledge that the longer the data time series the better the model performance. However, it is also noticeable that the minimum sum is 4 years in the Table, and this leads to a motivation of asking why, is it the case that the sum of at least 4 years data makes a critical point for good prediction, and, does the community structure need such a period of time to form a repeated cycle longer than one year; moreover, are these the case study specified or they have broader sense. All these are attractive issues for further investigation.

Table 6. Performance of k-NN approach for specially selected datasets.
The inputs were data for preceding years' of 2013 and the data of 2013 were testing data.

Subset	Number of years	Sensitivity (%)	Specificity (%)	Precision (%)	ACC (%)	F1
sub0018	4, 6, 7	71.43	100.00	100.00	83.33	0.8333
sub0027	4, 6, 7	85.71	100.00	100.00	91.67	0.9231
sub0065	4, 5, 7	85.71	100.00	100.00	91.67	0.9231
sub0089	6, 7	71.43	100.00	100.00	83.33	0.8333
sub0119	7	**100.00**	**100.00**	**100.00**	100.00	**1.00**
sub0181	7	**100.00**	**100.00**	**100.00**	100.00	**1.00**
sub0272	7	85.71	100.00	100.00	91.67	0.9231
sub0503	6	71.43	100.00	100.00	83.33	0.8333
sub1968	6, 7	85.71	100.00	100.00	91.67	0.9231
total set	7	85.71	100.00	100.00	91.67	0.9231

Figure 4 is an illustration of the indices in Table 6. Subsets sub0119 and sub0181, with their best performance, are accompanied by 27 and 18 variables (OTU features). This suggests that the dataset with small number of features after feature selection may yield better output than those with more variables (features).

Within the machine learning approaches, the performance appears to vary with the size of the sub dataset. In k-means clustering, the sub dataset sub1968 (having 1968 features or variables) shows its highest F1 score (Table 3), while in SVM approach, sub dataset sub0027 (contained 27 features or variables) exhibits the best performance among the 9 sub datasets in terms of the F1 score. For the k-NN output, the subset sub0089 and sub0181 show the best performance with their F1 scores above 0.92. The significance of the feature selection processing has been demonstrated here and it appears an advantageous practice to carry out feature selection on an original OTU dataset before going on further machine learning investigation. Whilst it is difficult to detect from the current dataset whether or not there exists some relation (e.g. the proper ratio of the selected features against the total number of the features) between the size of a sub dataset and the optimized performance of an algorithm working on it. It is interesting to observe and investigate this relation in more trails with different datasets.

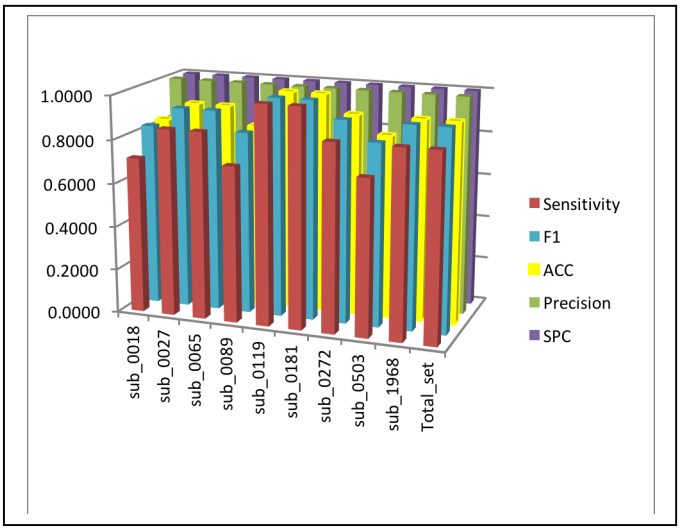

Figure 4 Comparison of indices yielded from k-NN approach for different combination

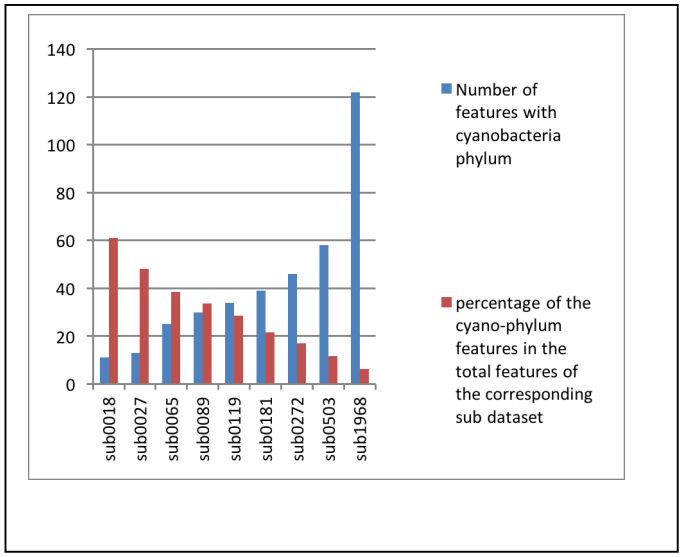

Figure 5 Number of features with Cyanobacteria phylum in the sub datasets and the percentage of these features in the total features of the corresponding sub dataset.

datasets in terms of the F1 score. For the k-NN output, the subset sub0089 and sub0181 show the best performance with their F1 scores above 0.92. The significance of the feature selection processing has been demonstrated here and it appears an advantageous practice to carry out feature selection on an original OTU dataset before going on further machine learning investigation. Whilst it is difficult to detect from the current dataset whether or not there exists some relation (e.g. the proper ratio of the selected features against the total number of the features) between the size of a sub dataset and the optimized performance of an algorithm working on it. It is interesting to observe and investigate this relation in more trails with different datasets.

Within the machine learning approaches, the performance appears to vary with the size of the sub dataset. In k-means clustering, the sub dataset sub1968 (having 1968 features or variables) shows its highest F1 score (Table 3), while in SVM approach, sub dataset sub0027 (contained 27 features or variables) exhibits the best performance among the 9 sub datasets in terms of the F1 score. For the k-NN output, the subset sub0089 and sub0181 show the best performance with their F1 scores above 0.92. The significance of the feature selection processing has been demonstrated here and it appears an advantageous practice to carry out feature selection on an original OTU dataset before going on further machine learning investigation. Whilst it is difficult to detect from the current dataset whether or not there exists some relation (e.g. the proper ratio of the selected features against the total number of the features) between the size of a sub dataset and the optimized performance of an algorithm working on it. It is interesting to observe and investigate this relation in more trails with different datasets.

The OTU table of the dataset shows that there are only 193 features with Cyanobacteria phylum and this is 4.75% of the total 4061 features. Fig. 5 shows the number of features with Cyanobacteria phylum in each sub dataset, and the corresponding percentage of the Cyanobacteria-phylum feature in each sub dataset. The percentage varies from 61.11% to 6.20%, from the sub dataset sub0018 to sub1968, with 18, 27, 65, 89, 119,181,272, 503, and 1968 features respectively. It appears that 1) good performance of the machine learning approach exists in low-percentage sub datasets to high-percentage sub datasets. For example, sub1968 (Table 3) has low-percentage (6.20%); sub0089 and sub0181 (Table 5) have medium percentages (33.71% and 21.55% respectively), and sub0018 and sub0027 (Table 4) have relatively high-percentages (61.11% and 48.15% respectively). This may suggest that, apart from the Cyanobacteria phylum, there might be other factors in action in the community structure in the bloom-forming dynamics and it is interesting to carry out further investigation on this line.

Tromas et al. randomly selected the training dataset in predicting bloom timing with symbolic regression (SR) [13]. In the study here we used the upfront portion of the total dataset

as the training set, highlighted the prediction function of the models, with satisfied performance.

In general, machine learning approaches in this case study showed good performance in the bloom prediction as evaluated by the indices; among the methods deployed, k-NN approach demonstrated superiority over the other two methods. Supported by the quick analysing facility with metagenomic techniques, the bloom prediction may be realized without much cost for collecting large amount of environmental data.

IV. Conclusion

With high values of evaluation indices for the prediction performance as good as 1.0 for the F1 score and 100% for the rest ones, this case study demonstrated the feasibility of using machine learning approaches in the Cyanobacteria bloom prediction with only metagenomic sequence data. Supervised and unsupervised machine learning methods, k-NN, SVM, and k-means clustering, all showed adequate performance, but k-NN appeared to surpass the other methods, in this case study. k-means clustering as an unsupervised method has its merit specifically in the practice of bloom prediction using metagenomic data, and it appears proper to attach importance to this type of approach.

In the k-NN approach, it was seen that longer data set led to better prediction performance (7 years in this case study). The critical number of the sum of preceding years for adequate performance of the bloom prediction (in this case study it appears to be 4 years) is worth further investigation.

Feature selection processing is significant in that it reduces the dimension of the dataset and save the modelling time whilst maintains or even increases the goodness of the performance of the modelling. As seen in the processing of 9 sub dataset obtained from the feature selection in this case study, the variables (features) contained in the subset were from 0.44% (sub0018) to 48.46% (sub1968) of the original amount of the variables (features), but the output of the modelling based on these subsets possessed similar degree of goodness in terms of those evaluation indices, or had even better results (Table 5 and Table 6). Meanwhile, it is noticeable that the two best output for the forecasting based on the 7years series are for the subset sub0119 and sub0181 which correspond 2.93% and 4.46% of the total number of the features (Table 2 and Table 6) respectively.

The machine learning approaches using merely the DNA sequence data showed better performance in this case study than the environmental factor driven models reported. Traditional environmental data driven models require acquisition of physical, chemical and biological data which involve considerable cost, from the field work equipment to the analysing instruments, and the analysis of samples can be time consuming. The metagenomic data based machine learning approaches are efficient, economic, and faster, possessing the advantage and potential for being adopted as a promising means in the bloom prediction practice.

Acknowledgements

The authors are grateful to Dr Nicolas Tromas for kindly providing the sequence data that were used in this research.

Reference

[1] Carmichael, W.W. and Boyer, G.L. "Health impacts from cyanobacteria harmful algae blooms: Implications for the North American Great Lakes." *Harmful algae* 54 (2016): 194-212.

[2] Le, C., Y. Zha, Y. Li, D. Sun, H. Lu, and B. Yin. "Eutrophication of lake waters in China: cost, causes, and control." *Environmental Management* 45, no. 4 (2010): 662-668

[3] Zurawell, R.W., Chen, H., Burke, J.M. and Prepas, E.E. "Hepatotoxic cyanobacteria: a review of the biological importance of microcystins in freshwater environments." *Journal of Toxicology and Environmental Health, Part B* 8, no. 1 (2005): 1-37.

[4] Wu, S.K., Xie, P., Liang, G.D., Wang, S.B. and Liang, X.M. "Relationships between microcystins and environmental parameters in 30 subtropical shallow lakes along the Yangtze River, China". *Freshwater Biology*, *51*(12), (2006): pp.2309-2319.

[5] Yabunaka, K.I., Hosomi, M. and Murakami, A. "Novel application of a back-propagation artificial neural network model formulated to predict algal bloom." *Water Science and Technology* 36, no. 5 (1997): 89-97.

[6] Wu, G. and Xu, Z. "Prediction of algal blooming using EFDC model: Case study in the Daoxiang Lake." *Ecological Modelling* 222, no. 6 (2011): 1245-1252.

[7] Vilán, J.V., Fernández, J.A., Nieto, P.G., Lasheras, F.S., de Cos Juez, F.J. and Muñiz, C.D. "Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain)." *Water resources management* 27, no. 9 (2013): 3457-3476.

[8] Li, W., Qin, B. and Zhu, G. "Forecasting short-term cyanobacterial blooms in Lake Taihu, China, using a coupled hydrodynamic–algal biomass model." *Ecohydrology* 7, no. 2 (2014): 794-802.

[9] Lou, I., Xie, Z., Ung, W.K. and Mok, K.M. "Integrating Support Vector Regression with Particle Swarm Optimization for numerical modeling for algal blooms of freshwater." *Applied Mathematical Modelling* 39, no. 19 (2015): 5907-5916.

[10] Pace, N.R., Stahl, D.A., Lane, D.J. and Olsen, G.J. "The analysis of natural microbial populations by ribosomal RNA sequences." In *Advances in microbial ecology*, pp. 1-55. Springer US, 1986., 9, 1-55.

[11] Winter, C., Hein, T., Kavka, G., Mach, R.L. and Farnleitner, A.H. "Longitudinal changes in the bacterial community composition of the Danube River: a whole-river approach." *Applied and environmental microbiology* 73, no. 2 (2007): 421-431.

[12] Tan, B., Ng, C., Nshimyimana, J.P., Loh, L.L., Gin, K.Y.H. and Thompson, J.R. "Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities." *Frontiers in microbiology* 6 (2015).

[13] Tromas, N., Fortin, N., Bedrani, L., Terrat, Y., Cardoso, P., Bird, D., Greer, C.W. and Shapiro, B.J. "Characterizing and predicting cyanobacterial blooms in an 8-year amplicon sequencing time-course." *bioRxiv* (2017): 058289.

[14] James, G., Witten, D., Hastie, T. and Tibshirani, R.. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

[15] Bermingham, M.L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., Agakov, F., Navarro, P. and Haley, C.S. "Application of high-dimensional feature selection: evaluation for genomic prediction in man." *Scientific reports* 5 (2015).

[16] Kira, K. and Rendell, L.A. "The feature selection problem: Traditional methods and a new algorithm." In *Aaai*, vol. 2, pp. 129-134. 1992.

[17] Kononenko, I., Šimec, E. and Robnik-Šikonja, M. KKononenko, I., Šimec, E. and Robnik-Šikonja, M. "Overcoming the myopia of inductive learning algorithms with RELIEFF." *Applied Intelligence* 7, no. 1 (1997): 39-55.

[18] Robnik-Šikonja, Marko, and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF." *Machine learning* 53, no. 1-2 (2003): 23-69.

[19] Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression." *The American Statistician* 46, no. 3 (1992): 175-185.

[20] Witten, Ian H., et al. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[21] Steinwart, Ingo, and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[22] Campbell, Colin, and Yiming Ying. "Learning with Support Vector Machines, 2011, Morgan and Claypool." *ISBN* 1967923169: 13.

[23] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, no. 1 (1979): 100-10