Morgan Tholl

CE 510: Machine Learning Methods in Civil Engineering

Using Machine Learning Methods to Predict Chlorophyll Concentration in the Willamette River

Introduction

Modeling chlorophyll in waterbodies is an important and complex problem in environmental engineering. It is indicative of the algae content, which can have a large impact on the health of the system. Excessive algae growth causes nutrient depletion in the system, blocks sunlight to other underwater plants, and can release dangerous toxins into the system. The eventual die-off when the algae exceeds the capacity of the system leads to dangerous fluctuations in the dissolved oxygen levels, which can cause further die-offs of marine plants and animals. This is the cause of famous red tides in the Gulf of Mexico or off the Oregon Coast, causing shellfish to become inedible. It's a growing problem for drinking water agencies, who must cope with increasing amounts of eutrophication in water sources due to human impacts and climate change. It can cause smell or odor issues, or even render water non-potable.

There are useful theoretical methods to estimate algae content, such as the use of a limiting nutrient model. However, the real-life mechanics are very complex. The mechanics depend on site-specific variables such as water use, watershed characteristics, meteorological conditions, natural daily and seasonal fluctuations, ecological impacts, and other variables that are difficult or impossible to incorporate in a theoretical model.

The complexity of the problem and the potential for hidden higher-level variables makes this problem a good candidate for machine learning using a neural network, given sufficient environmental data.

Dataset and Features

Machine learning methods require large amounts of data to find hidden patterns, so a suitable location for this study would have a significant amount of data available. The USGS gage for the Willamette River underneath the Morrison Bridge has a large amount of available data online, including in-situ monitoring of chlorophyll content for over 2 years. There is National Weather Service data available that can be used for meteorological data in this location as well. Since in-situ total chlorophyll modeling was performed 01/22/2009 – 09/29/2011, this is the date range that will be used for the rest of the data. The raw number of instances is 981. The data was sanitized prior to use, which reduced the number of instances to 940 due to missing data. Table 1 shows the features, which consist of six input features and the target.

Tidally filtered discharge is discharge data that has been manipulated to exclude the effects of tidal fluctuation, since the USGS gage is close to the mouth of the river and its water levels are impacted by the tide. Specific conductance is an easy way to measure the ionic content of the water. It is often used as a water quality parameter as a proxy for how much other particulate matter and dissolved content is in the water.

Based on theoretical models of algae growth, we should expect total chlorophyll to be positively correlated with water temperature and air temperature. Since discharge is lower in the summer, total chlorophyll will likely be negatively correlated with discharge. Precipitation has competing effects. Lower precipitation is associated with summertime conditions, but rain events wash nutrients into the waterway, which cause algae growth. Oxygen in the system is even more complex and there are many competing effects such as daytime oxygen production, nighttime consumption, oxygen consumption upon death, shading of underwater plants, and more.

Methodology

Total chlorophyll will be used as the target value in a regression problem involving the set of scalars listed in Table 1. The data will be standardized and split into three sets: 80% of the data in a training set, 10% in a validation set, and 10% in a testing set. The training set will be used to build the model. The validation set will be used for model selection, and the recommended final model will be presented using statistics from the reserved testing set.

Three machine learning methods will be used to build a regression model: support vector machines, artificial neural networks and random forests. For each method, a grid search will be used to test models with parameter options. The model with the best R^2, while preventing overfitting, will be selected. Overfitting will be defined as a model that performs significantly better on the training data than the validation data.

A grid search will be used to optimize the support vector model. The parameters to tweak are the kernel type and epsilon value for penalties. The model performance on the training set and validation set can be compared to determine whether the model is overfitting.

Artificial neural networks can be implemented many ways, so a simple grid search will not be sufficient for model selection. Different optimizer methods, as well as techniques such as regularization, batch normalization, early stopping, and dropout, will be tested to determine which model gives the lowest mean squared error while not overfitting the training data. After evaluating these techniques, the optimal model structure will be selected and a grid search will be used to tune the hyperparameters (number of hidden layers, number of neurons, learning rate, and optimizer-specific values) to obtain the best R^2 value.

A grid search will also be used to find the optimal parameters for a random forest model. A grid search will be performed for random forest regression models. The search will test options for the number of estimators, maximum depth, and maximum number of features, to find the model with the best performance. Bootstrapping will be used (limiting the number of samples in each iteration) along with a minimum sample split size to avoid overfitting.

The R^2 and mean squared error (MSE) values of the best model from each method will be presented.

Results and Discussion

The optimal support vector machine model used a polynomial kernel function, an epsilon value of 0.1 and a 'C' regularization term of 10. The R^2 of the training set was 0.83 and the R^2 of the validation set was 0.85, indicating that the model is fit fairly well and is not overfitting.

The artificial neural network used an Adam optimizer with early stopping and batch normalization to prevent overfitting. Adding in dropout did not have a significant effect on the model error. Adding in batch normalization and regularization increased the model error, but without these methods, the model tended to overfit the training data. Batch normalization was added to the model because it contributed less error. Once the model structure was selected, the hyperparameters found using the grid search were 2 hidden layers with 12 neurons each and a learning rate of 1E-3. The R^2 of the training set was 0.79 and the R^2 of the validation set was 0.77, so overfitting is not likely an issue with this model. In general, it was difficult to avoid overfitting. The model was simplified from earlier versions in order to avoid overfitting.

A simple decision tree was able to achieve a validation set R^2 of 0.75 and a training set R^2 of 0.74 using a minimum sample split of 20. The random forest method had a serious overfitting problem. Some models achieved an R^2 of 0.9, but would not generalize for the validation set. The optimal random forest model, found using a grid search for optimal parameters, had a maximum depth of 3, a maximum number of features equal to the square root of the number of features. and 300 estimators. Bootstrapping was used with a maximum of 200 samples per iteration, and a minimum sample split size of 20. The random forest model achieved a training set R^2 of 0.66 and a validation set R^2 of 0.57. A better model could not be achieved without overfitting, so this method is likely a poor strategy for predicting total chlorophyll.

A summary of the optimal performance from each model is shown in Table 2. The performance metrics presented are based on the test set.

On both performance measures, the support vector machine model performed the best, followed by the artificial neural network and then the random forest model. It's interesting to note that for all three models, the R^2 of the test set was actually higher than the training or validation R^2 values, despite being reserved from training and model selection.

The feature importance was also investigated. The relative importance of features in the random forest model using an impurity-based method are listed in Table 3.

The relative importance shows that the most important features for the random forest model were tidally filtered discharge, water temperature, and air temperature. Based on these features, predicting seasonality was likely a key factor in predicting the total chlorophyll levels. Precipitation was almost entirely ignored in the model. A subset of the test set was created to investigate how the models performed on days with high total precipitation. On this "high precipitation" test set with only 17 instances, the support vector machine achieved an R^2 of 0.50, the neural network achieved an R^2 of 0.29, and the random forest achieved an R^2 of 0.73. While the random forest had the lowest performance measures overall, it performed best at generalizing instances with high precipitation.

The permutation feature importance for the best-performing models are shown in Table 4.

Compared to the other methods, random forest put more weight on the air temperature. Data visualization shows that the water and air temperatures are essentially the same, and after standardization they could be combined into one “temperature” feature. Otherwise, temperature may have an over-sized effect on the prediction. Dissolved oxygen is barely a factor in the random forest model, but it is significant in the models that perform better.

Surprisingly, precipitation and specific conductance were not significant in any of the models, and it's possible that they could be excluded for better results.

Conclusions and Recommendations for Future Work

The best-performing model was the support vector machine model, with an R^2 of 0.87 and a mean squared error of 0.40 ug/L, which is very high performance considering that most existing chlorophyll models almost always include nutrient availability. Perhaps when reliable data on nutrient availability is not present, a location-based machine learning model like these could forecast the expected chlorophyll content using other water quality parameters.

None of the tested models had high feature importance for precipitation, which is surprising since rain events wash nutrients into the system. Most theoretical models are based on availability of these nutrients. Perhaps the model was confused by the competing effect of lower precipitation in summertime conditions, or by the delayed response of the system as algae grows after a rain event. Removing certain features, or testing the model's accuracy on and after rainy days, would provide more insight into the model's capabilities. A time-series method would be more relevant for this application, due to the dependency of algae growth on its current concentration (first-order kinetics).

Continuing to monitor chlorophyll in-situ in the Willamette River would provide valuable data that could be used to train a better model. Since the model only included 2.7 years of data, the model is unlikely to generalize well for yearly weather fluctuations or extrapolate in the face of progressive climate change. The model would generalize better with more data. More data would mean that a model could be built using only summertime conditions, when an algae bloom is likely. I recommend continued in-situ chlorophyll monitoring in the Willamette River.

Table 1. Data for total chlorophyll modeling at the Willamette River in Portland, OR. All values are daily averages.

Data	Units	Source	Use
Total Chlorophyll	µg/L	USGS	Target Value
Water Temperature	°C	USGS	Input Feature
Tidally Filtered Discharge	cfs	USGS	Input Feature
Dissolved Oxygen	mg/L	USGS	Input Feature
Specific Conductance	uS/cm	USGS	Input Feature
Air Temperature	°F	NWS	Input Feature
Total Daily Precipitation	In	NWS	Input Feature

USGS:
USGS Surface-Water Daily Statistics for the Nation: USGS 14211720 Willamette River at Portland, OR
https://waterdata.usgs.gov/nwis/dvstat?referred_module=sw&search_site_no=14211720&format=sites_selection_links
NWS:
National Weather Service Forecast Office: Climatological Data for Portland Area, OR
https://w2.weather.gov/climate/xmacis.php?wfo=pqr

Table 2. Summary of the performance measures from the optimal models found using each machine learning method

Method	R^2 of test set	Mean Squared Error of test set
Support Vector Machine	0.87	0.40 ug/L
Artificial Neural Network	0.85	0.46 ug/L
Random Forest	0.76	0.74 ug/L

Table 3. Relative importance of features in the random forest model using an impurity-based method

Feature	Relative Importance
Tidally Filtered Discharge	0.273
Dissolved Oxygen	0.100
Water Temperature	0.353
Air Temperature	0.224
Precipitation	0.002
Specific Conductance	0.048

Table 4. Average feature importance using a permutation method on the best performing models from each method

	Tidally Filtered Discharge	**Dissolved Oxygen**	**Water Temp**	**Air Temp**	**Precipitation**	**Specific Conductance**
Support Vector Machine Importance	0.168	0.435	0.291	0.046	0.013	0.046
Artificial Neural Network Importance	0.058	0.596	0.316	0.019	0.002	0.010
Random Forest Importance	0.256	0.048	0.471	0.195	0.001	0.030