Project Report
ISE 244
Xudong Guo

1. Introduction

Water is one of the essential elements in our daily life. Government has an important responsibility to identify safe water and ensure good quality water delivered to citizens. There are lots of measures and variables when it comes to classifying safe water.

Water Quality Index is one of the important tasks to identify safe water. Its calculation usually requires several variables and complex steps. A lot of sub-indices are calculated during the process. This can cause potential errors during manual calculation. There are also many different WQI equations all over the world. In general, the manual calculation of WQI requires a lot of time, involves complex processes, and can have inconsistent results based on different equations. As a result, machine learning can be useful to solve these issues. AI models can predict the WQI with given variables without the need to go through the complex calculation.

The solution proposed by the research paper I found is to use novel hybrid models to improve the accuracy and performance over the basic models. The hybrid models used in the paper are bagging, cross validation parameter selection, and randomized filtered classifier. These are integrated with basic models including random forest, random tree, M5P and reduced error pruning tree. The paper also used correlation to construct input combinations from the ten features.

After studying the above research paper, this project is set to use the similar approach to solve a water quality classification problem. A Kaggle water quality classification data set is chosen. This project aims to compare the performance of different classification models.

2. Objective / Methodology

The project aims to compare the classification models including logistic regression, random forest, and cost complexity pruning along with hybrid models, such as bagging, cross validation.

2.1 Data Collection and Preprocessing

Data set is downloaded from Kaggle.com. The data contains 3276 rows, with 9 features. Potability category is labeled either as 0 or 1, representing if the water body is safe for humans to consume. Basic exploratory data analysis is performed to check null values, correlations, and category distribution. The null values in each feature are replaced with the median value, since this is a small data set. The data is then normalized to a scale of 0 to 1.

The data is then split into training and testing sets with a 7:3 ratio.

2.2 Input Combination

Best input combination is needed for each model. The correlation coefficient between the nine features and potability is computed. Then the feature with the highest coefficient was first considered as the input. Then each of the following features was added to the previous input combination, to a total of nine combinations. The ranking of correlation coefficient is: Solids, Organic carbon, Chloramines, Sulfate, Hardness, Conductivity, Trihalomethanes, pH, Turbidity. All the input combinations are tested. Accuracy and precision are the metrics to determine the best combination.

2.3 Models

Sklearn library is used for all the models.

For this classification problem, two basic models are used: logistic regression and random forest. Logistic regression is a basic model that serves as a ground performance for this problem. For random forest, there are balanced mode for uneven class distribution and cost complexity pruning that prevents overfitting. These can be enabled with a parameter input when constructing the model.

On top of the basic models, grid search is being used to find the optimum parameters. Bagging is used as a meta hybrid method. It is used to compare with the basic models in order to see if there is any performance improvement. Cross validation is also used to build and improve the fit of the model. All the models are optimized with 10-fold cross validation.

2.4 Training result and analysis

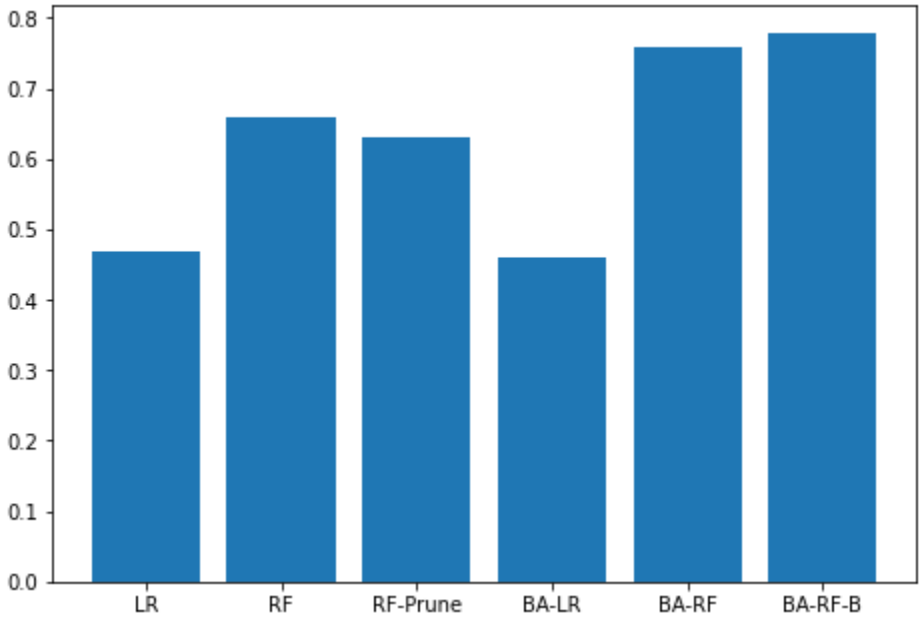
Below is the table of accuracy for each model in the training and testing. The yellow marked cells are the highest accuracy of each model on the testing set.

		1	2	3	4	5	6	7	8	9
Logistic Regression	training	0.49	0.52	0.51	0.52	0.52	0.51	0.51	0.51	0.52
	testing	0.5	0.54	0.55	0.54	0.54	0.54	0.54	0.53	0.53
Random Forest	training	0.72	0.84	0.85	0.86	0.87	0.87	0.88	0.98	0.99
	testing	0.55	0.56	0.55	0.61	0.63	0.62	0.61	0.64	0.63
Cost Complexity Pruning	training	0.37	0.61	0.88	0.63	0.97	0.86	0.98	0.98	0.98
	testing	0.42	0.57	0.57	0.6	0.63	0.62	0.61	0.65	0.64
Bagging with Logistic Regression balanced class	training	0.53	0.47	0.49	0.5	0.49	0.49	0.51	0.51	0.5
	testing	0.54	0.52	0.51	0.51	0.51	0.52	0.52	0.52	0.5
Bagging with Random Forest	training	0.72	0.75	0.77	0.78	0.89	0.92	0.93	0.99	0.93
	testing	0.55	0.56	0.57	0.61	0.63	0.63	0.62	0.66	0.65
Bagging with Random Forest balanced class	training	0.72	0.75	0.76	0.78	0.9	0.9	0.92	0.99	0.99
	testing	0.55	0.56	0.57	0.6	0.62	0.62	0.63	0.65	0.65

Logistic regression performed poorly with low accuracy on both training and testing sets. Random forest fit well on the training set, but got low accuracy on the testing sets. Overall, combination 8 performed best among all the inputs. This combination is where ph is added.

While cross validation is applied during the training phase, the models still have low accuracy on the testing sets. This might be because the data size is too small. There weren't enough samples to represent the whole population. Based on this assumption, a bigger data size could possibly increase the testing accuracy.

With precision included, a more interesting pattern could be seen. Below is a bar chart that shows the best precision for predicting positive potability in each model.



Precision is noticeably increasing with bagging added on top of random forest, and even more with balanced mode enabled. This means that the model is actually improving with the hybrid models despite the accuracy staying low.

There are other metrics such as recall and F1-score being tracked during the training phases. There are no obvious patterns observed. Recall stayed relatively high for the negative cases. This is reasonable since the class distribution is unbalanced.

The final model is trained on the whole data set. The input combination is the 8th combination. The model is chosen to be bagging integrated with random forest in balanced mode. The final accuracy is 0.98 with a precision of positive potability to be 1. This proves that the model is capable of accurately classifying water quality. A better data set should improve the training and testing accuracy.

3. Reflection

There is definitely more that can be done to improve the current state of the model. One interesting finding is that the input combination 8 is standing out among the other inputs. This can indicate that pH is a significant feature affecting the potability, even though their correlation is low. To check the range and variation of features, it is obvious that pH has a small range and variation. But the ph value, in theory, determines

if water is acidic or not; a neutral pH value usually means the water is at least not harmful. This implies that correlation might be deceiving in the process of machine learning. More input combinations should further be tested to prove this point.

A bigger data set should be used in order to improve and validate the model. In the current state, the model suffers from low accuracy on testing sets, which is not promising for the model to be deployed for any service.

In this project, both accuracy and precision are significant metrics, especially precision for positive cases. A model serving for such cases should ensure the precision of positive cases to be as close to hundred percent. This might cause the recall to be low, which could potentially cause a waste of resources when classifying good quality water to be bad. However, this can always be solved by further investigation and manual adjustment. To train an appropriate model for water quality, features must be studied closely to identify the significant features.

4. Conclusion

Water quality is important to human health. Even though machine learning can be a convenient tool to predict water quality, its result requires more attention and analysis before getting employed to related facilities. While this project proves that classification on water quality can potentially be precise and accurate, the final model is far from ready to be published. More data and features should be studied to ensure that the model is trustworthy.