A Machine Learning Approach for Real time data analysis

Nazmul Hasan
ID: 1330824642
Section: 01

Jahidul Islam Munna
ID: 1420474042
Section: 01

Mahamud Hasan
ID: 1420453042
Section: 01

A.M. Almarufuzzaman
ID: 1420469042
Section: 01

***Abstract*—We use water regularly in our everyday life. However, sometimes we use pollute and dirty water due to negligence. To avoid using pollute and dirty water we made a portable device which will measure water quality by using pH, EC, Turbidity and Temperature sensors and then sent those data to Firebase to calculate and determine water quality. We have also successfully implemented machine-learning algorithm for early prediction about tomorrow's water quality so that user can get ready. Nave Bayes, kNN and Learning Regression were used as learning algorithms in the project. Nave Bayes is very fast and gives great accuracy. kNN is a very straightforward algorithm. Learning Regression was used for generating pH, EC and other features for a day of the future. C++ is used as programming language to implement those machine learning algorithms. kNN gives approximately 94% accuracy.**

Keyword: Machine Learning, Linear Regression, Sensor, Real time, kNN, Nave Bayes, Accuracy

I. INTRODUCTION

We made a portable device with Node MCU and four different Sensors to get real time data from the sensors. The device will measure water temperature, pH level, electrical conductivity (EC), and turbidity and then it will pass those data to firebase. By comparing those data with the standard data stored in server, the server will let the user know whether the water is safe to use or not. In the server end, there will be used a machine-learning algorithm to predict the water quality for future use and a notification system to let the user know about the water quality.

Undoubtedly, water is the most important thing in our life. Every day we use water for various cause but actually, we only care about our drinking water. We usually neither check nor care about our reservoir water quality level. However, most of the work of our household or industry depends on reservoir water. The reservoir does not polluted in a day; it takes some time to be become polluted. However, it is time-consuming and expensive to check reservoir water quality every day. We usually forget to check reservoir water quality very often and only get concerned when someone get affected due to water pollution or a dirty water flow on the main waterline but then that is too late. Because in the meantime we have already used the unsafe water.

In our project, we are trying to implement such a system that can give the real-time water quality check as well as the early prediction about the usability of water. For this early prediction, we are going to use machine-learning algorithms to train predictor from real time water quality data.

II. BACKGROUND

Water quality depends on some physicochemical component measurement. Usable water quality measurement depends on the water-using task and that task indicates about how the measurement we need to achieve. We are working on the reservoir water, which is allocated to our daily usage. To determine reservoir water usability, we need to check pH, Electrical Conductivity (EC), Turbidity and Temperature. Therefore, we used four sensors to collect the data on water quality. pH sensor checks if the water is close to acid or base. The standard safe pH level for water is 6.5-8.5. EC sensor measures the conductivity level of water. It determines how much salt the water has in it. The safe EC level is below 1.5. Turbidity sensor measures the cloudiness of water. The standard value of turbidity is 10 NTU. The last important sensor is the temperature, which measures the temperature level of water. The standard value of water is 20-30 C.

We take the data from the sensor and update those to the Firebase. For our early prediction of reservoir water usability, we are going to use a machine learning approach. It is necessary to know, whether the water is safe to use or not. We read study some related work, which are close to our work for solving our project.

A work named Predicting Sleep Using Consumer Wearable Sensing Devices was done, where they determined whether the user is in sleep or awake. At first, they preprocessed data, feature scaling and try five algorithms to predict. They used binary and multi-class classification. Logistic regression, kNN, SoftMax Regression; Support Vector Machine is used for binary and multi-class classification and kNN.

Another work name Smart Bin: An Intelligent Waste Alert and Prediction System Using Machine Learning Approach is very close to our project where they use real-time data and analyze the data they predict future data and send a notification to authority. They used Microsoft Azure machine learning studio for graph and using the pick of graph, they update the authority.

III. DATA AND DATA SOURCE

The data we have used came from different sensors of the device that we have made for this project. Temperature, PH, Turbidity, and Electrical Conductivity (EC) sensors are mainly used to get those data. All of these four attributes are the primary features to determine whether the water is safe to use

or not. Temperature gives us the information if the water is too hot or cold. We want to use neither hot nor cold water usually. So, determination of temperature is necessary. We have used values greater than 15 C and less than 45 C to be normal. Any higher or lower temperature is considered to be unsafe to use. PH, an important feature of water. It detects the acidity of it. The range of pH level is 0 to 14 where 7 indicates a neutral substance. Usable water has a range of 6 to 8.5. Any lower value from 6 or higher from 8.5 is considered to be an acid or alkaline respectively. We also need to know if the water is opaque or not. We simply do not want to use any dirty water generally. In order to detect dirty or clean water, we used a turbidity sensor here. It gives a value from 0 to 11. Where value less than 8 is dirty and greater than or equal to 8 is considered to be clean water. The last feature is Electrical Conductivity (EC) of water. It detects the capability of water to pass electrical flow. It is related to the concentration of ions in it. Mainly EC sensor detects whether the water is salty or not. Seawater has a very high range of conductivity. Safe water has a range of conductivity from 0 to 1.5 ms/cm. We put our device in a various solution of water and took the reading. We changed the water condition by reducing temperature, adding salt or detergent, mixing acid with it and so on. We took all of the data in our dataset. In addition, we took the date and time as well. We organized all the data into 6 columns. We put Date, pH, Temperature, Turbidity, and EC values in a separate column. The last column is the output column also known as the label column.

A Sample data chart is given below:

Date	pH	Temp	Turbidity	EC	Safe
07-07-18	6.408	31.25	8.5	0.1155	YES
07-07-18	7.2527	29.25	8.8	0.1105	YES
07-07-18	7.125	29.25	9	0.137	YES
09-07-18	7.1819	29.063	9	1.2419	YES
09-07-18	5.0233	31.313	8.3	0.9626	NO
12-07-18	6.1576	31.25	8.3	0.9668	YES
12-07-18	6.087	31.063	8.3	0.9796	YES
12-07-18	5.7729	30.938	6.3	0.9881	NO
16-07-18	6.0756	30.5	8.3	0.6135	YES
16-07-18	8.9481	30.313	8.3	3.1243	NO

Figure: A sample Dataset

IV. METHODOLOGY

For our particular project, we have used our own dataset, which we have acquired from our Node MCU based portable device to train the learning algorithms for early prediction of safe or unsafe reservoir water. We used four features (pH, Temperature, Turbidity and EC) to determine if the reservoir water is safe to use or not. Safe range for particular features are - 15° C < Temperature < 45° C; 6.0 ≤ pH ≤ 8.5; turbidity ≥ 8.0; 0 $ms\ cm^{-1}$ ≤ EC ≤ 1.5 $ms\ cm^{-1}$.

We have used Naïve Bayes, K-Nearest Neighbor (kNN) and Learning Regression as our learning algorithms. Those are faster and give great accuracy for few amount of features. Since, our dataset contains only four features (pH, Temperature, Turbidity and EC), those learning algorithm will work nicely. Our initial accuracy estimation for Naïve Bayes is approximately 88%, for kNN is 94% and for Linear Regression is 84%.

We have used C++ as our programming language to implement all of the algorithms. We are giving our dataset as text file, which we are getting from Firebase by using JavaScript.

We relied on various Paper, class lectures and online sources for understanding and planning to complete our project. Class lectures were very important for understanding and implementing algorithms.

V. ALGORITHM CLASSIFICATION

Our problem was how we can classify water by looking at the four features. We have four feature, around five hundreds of dataset and two classes. After conducting some research, we choose kNN and Naïve Bayes classifier algorithm. As we are trying to predict tomorrow's water quality, so we don't have the feature data for the next morning. Here we applied Linear Regression to solve this problem.

Naive Bayes: We used Naïve Bayes classification because of its speed and accuracy. It works on some mathematical theory, so there is no complex training method needed. It's a member of probabilistic classifier.

Our problem was classify the water from the four features. As we know, it works on probabilistic theory, so we had to summarize data for faster Gaussian Probability calculation. We created two different 2D array to store data. First array to save the only safe water training data and the second one to save unsafe water data. Then we calculated the mean and Variance using these formula.

$$Mean, \quad \mu = \frac{1}{N}\sum_{k=1}^{n} X_k$$

$$Variance = (\delta)^2 = \frac{1}{N-1}\sum_{k=1}^{n}(X_k - \mu)^2$$

Then we find the posterior for both of the safe and unsafe class.

$$Posterior(safe) = \frac{P(safe) * P(temperature) * P(ph) * P(turbidity) * P(EC)}{evidence}$$

$$where, \quad P(x) = probabilty\ for\ x.$$

$$all\ the\ fetuare\ probabilty\ only\ for\ safe\ water$$

$$Posterior(unsafe) = \frac{P(unsafe) * P(temperature) * P(ph) * P(turbidity) * P(EC)}{evidence}$$

$$where, \quad P(x) = probabilty\ for\ x.$$

$$all\ the\ fetuare\ probabilty\ only\ for\ unsafe\ water$$

$$P(safe) = P(unsafe) = 0.5$$

$$P(feature) = \frac{1}{\sqrt{2\pi\delta^2}}\exp(\frac{-(x-\mu)^2}{2\delta^2})$$

$$evidence = P(safe) * P(temperature) * P(ph) * P(turbidity) * P(EC) + P(unsafe) * P(temperature2) * P(ph2) * P(turbidity2) * P(EC2)$$

where, ph indicates the temperature for safe water and $ph2$ for unsafe water.

After these calculations we get $Posterior(safe)$ and $Posterior(unsafe)$. If $Posterior(safe)$ is greater than $Posterior(unsafe)$ then we classify it as Safe water, otherwise it is Unsafe.

kNN: K-Nearest Neighbor, broadly known as kNN. It is a classification algorithm that classifies by checking the Euclidian distance from test to trains. This one is very straightforward algorithm.

First of all, we split the dataset into two different parts with the ratio of 70:30. 70% data for training and rest of the 30% data for testing.

For each of test instance, we find the Euclidian distance by all the features the store the distance in a vector. After finding distances from all the training instances, we sort the distances and consider closest k neighbors. Then depending on the majority, it decides which class the test data belong to.

This algorithm uses very simple calculation, so it is an easy task for a normal processor and it is very fast.

$$Euclidian\ Distance\ d(a,b) = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 + \cdots + (a_n-b_n)^2}$$

Linear Regression: Though our problem was a classification algorithm, but it is also possible to break it into parts, which is individually a regression problem, and we can then combine those results and classify it. We have the safe range for every feature.

We used linear regression to find individual value for pH, Turbidity etc. Then check if it is into the safe range or not.

Linear regression takes the training dataset and sends it to the learning algorithm and learning algorithm returns an equation.

X	Y
1	31
2	29
3	29

Table: Sample table for temperature.

X is the input index that indicates the day of the year. And Y indicates the temperature on X'th day. After passing the training dataset, we get an equation that is similar to

$$Y = A + BX$$

Then we can find the Y for any X. After finding the expected temperature, PH, Turbidity and EC, we check if it is the safe range or not. Then we declares it is safe or unsafe.

VI. RESULTS

We used three algorithms for our dataset and the accuracy we have achieved is satisfying. We have got -

Linear Regression: 84%
Naïve Bayes: 88%
kNN: 94%

The percentage for kNN vary for different value of k. We get almost 94% accuracy in kNN for k value 5.

VII. CONCLUSION

Water is a vital mineral for our everyday life. However, sometimes due to negligence we use pollute and dirty water and then become concern for few days. To avoid using pollute and dirty water we made a portable device which will measure water quality by using pH, EC, Turbidity and Temperature sensors. We have also successfully implemented machine learning algorithm for early prediction about tomorrow's water quality so that user can get ready. Since our problem was to see, how we can classify water by looking at the four features makes our problem as classifier problem. Nave Bayes, kNN is great classifier algorithm. They work nicely with few features and gives very accurate result. In addition, we used Linear Regression algorithm, which predicts feature data for a specific day in the future. At the end kNN gives us 94% accuracy for our dataset.

ACKNOWLEDGEMENT

We especially want to thank Mirza Mohammad Lutfe Elahi sir and Silvia Ahmed madam for giving us the idea of the project and give us the chance to work on this project.

REFERENCES

[1] Jessica Moore, Binghai Ling, "Human Activity Recognition using Smartphone Sensors",December 2016
[2] Jessica Moorea and Binghai Lingb,"Human Activity Recognition using Smartphone Sensors",16 December 2016
[3] B. Kotsiantis, Sotiris, I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." ,3-24, 2004.
[4] G. Chandrashekar and F. Sahin, A survey on feature selection methods, Comput. Electr. Eng., vol. 40, no. 1, pp. 1628, 2014.
[5] Cyril Joe Baby, Harvir Singh, Archit Srivastava, Ritwik Dhawan and P. Mahalakshmi, "Smart Bin: An Intelligent Waste Alert and Prediction System Using Machine Learning Approach",IEEE WiSPNET 2017 conference
[6] Miguel A. Garcia,"Predicting Sleep Using ConsumerWearable Sensing Devices",Stanford University project,2017