Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020)
IEEE Conference Record # 48766; IEEE Xplore ISBN: 978-1-7281-5371-1

Comparison of Water Quality Classification Models using Machine Learning

Neha Radhakrishnan
Department of Electrical and Electronics Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
rneha2096@gmail.com

Anju S Pillai
Department of Electrical and Electronics Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
s_anju@cb.amrita.edu

***Abstract*—Water resources are often polluted by human intervention. Water pollution can be defined in terms of its quality which is determined by various features like pH, turbidity, electrical conductivity dissolved oxygen (DO), nitrate, temperature and biochemical oxygen demand (BOD). This paper presents a comparison of water quality classification models employing machine learning algorithms viz., SVM, Decision Tree and Naïve Bayes. The features considered for determining the water quality are: pH, DO, BOD and electrical conductivity. The classification models are trained based on the weighted arithmetic water quality index (WAWQI) calculated. After assessing the obtained results, the decision tree algorithm was found to be a better classification model with an accuracy of 98.50%.**

Keywords—classification model; decision tree; support vector machine; naïve bayes; water quality index;

I. INTRODUCTION

The water resources of India play an indispensable role in the lives of people in India. The major water resources are rivers, lakes and ponds which provide adequate water for irrigation, electricity, transportation as well as for domestic purposes. To fulfill these needs, it has to be clean and free from contamination. But the recent trends have shown that the quality of water has been dropping to a very large extent and this has led to a large amount of water unfit for utilization.

By the World Health Organization (WHO) records, water-related challenges such as deprival of safe and clean water for domestic purposes, increasing urban pollution and the scarcity of water are growing at an alarming rate [1]. Hence, the decade 2018-2028 is announced as International Decade for Action, "Water for Sustainable Development" by the United Nations General Assembly [2]. The hike in the population rate is another factor for the scarcity of water. The quality of water is controlled by its certain constituent parameters, so when the effluents are unloaded into the water the concentration of these water parameters change which results in the decrease of water quality.

Over the past few years, a lot of investigative efforts have been done in the scrutiny of water quality and a variety of water quality models have been originated. Conventional methods include the collection of data manually and its statistical evaluation. M. A. Tirabassi et al. developed a statistical model to foretell the water quality without any reference to the chemical, biological, and physical relations [3]. In this paper the concept of the black box is used, that is with a known input a relatively dependable output can be predicted. According to Gaganjot Kaur Kang et al., when there is an availability of a large amount of data, big data analytics can be applied [4]. A challenge faced in this method is the accuracy of the water quality evaluation and prediction model. H. C. Guo et al. proposed a stochastic water-quality prediction systemthat was established to reveal the hazard characteristics of several attributes, based on Kalman-filtering and self-adaptive techniques. The system predicted the levels of BOD and DO of the Yilou River [5].

The quality of water is determined by various levels of different parameters. Amit Sinha et al. proposed a fuzzy model that inputs three parameters (conductivity, pH and hardness) and the model is then simulated using MATLAB[6]. Dataset is generated from several water samples gathered from different parts of Uttar Pradesh. Artificial neural network (ANN) algorithms have been largely used for the prognostication of water quality, one such example of an algorithm is the Back Propagation (BP) algorithm. An issue in this BP algorithm is the low accuracy percentage, so an improved artificial bee colony (IABC) algorithm has been presented. Comparing both the algorithms, there has been a surge of 25% more accuracy in the model. In this algorithm, the connection weight values between network layers and the threshold values of each layer are previously enhanced. Amir Hamzeh Haghiabi et al. proposed a learning that compares the performance of group method of data handling (GMDH), ANN, and SVM for forecasting water quality of Tireh River of southwest Iran [9]. Analyzing the results of the three algorithms, SVM model has a better performance when compared to accuracy. Wang Xuan et al. suggested a proposal to resolve classification, prediction challenge of non-linearity and inadequate data using the SVM. But, the practicality of SVM is affected due to the difficulty in choosing the suitable SVM parameters. This paper exhibits a fusion of SVM and particle swarm optimization to decide SVM free parameters for better accuracy of the model [7]. Salisu Yusuf Muhammad et al. suggested a suitable classification model based on machine learning techniques [8].

A comparison of five classification algorithms such as Naïve Bayes, K star, Bagging, J48 and Conjunctive rule has been done to find the important factors that assisted in classifying water quality of Kinta River, Perak Malaysia. Out of the five models, the Lazy model using the K Star algorithm was found out to be the best algorithm with 86.67%.of accuracy.

The following paper is organized into 4 sections. Section II presents the materials and the methods used for analysis of the study. The results and discussion of the experiment are presented in section III and the paper is concluded in section IV.

II. MATERIALS AND METHOD OF ANALYSIS

A. Experimental Datasets

Two datasets have been considered for testing of the water quality in the proposed work. The first dataset consists of twenty-eight different water quality parameters or features of the Narmada River, which flows through the state of Madhya Pradesh, for the year 2017-2018[14]. The values are sampled each month from different stations of the river. Few examples of the features are pH, chloride, BOD, potassium, nitrate, DO, electrical conductivity etc. The second dataset is the combined data for the historical water quality of certain locations in India. The parameter values in each column are the values over the years from 2003 to 2014 [15]. The dataset constitutes eight parameters such as total coliform, temperature, conductivity, faecal coliform, BOD, DO, pH and nitrate. There are around 1991 values of rows out of which some are invalid values which are eliminated during the process. Both the datasets are provided by the Indian government websites.

B. Water Quality Parameters

Water quality depends on the physical, chemical and biological factors of the water content. The change in the values of the parameters such as turbidity, temperature, biological oxygen demand, dissolved oxygen, electrical conductivity, nitrate and pH leads to a change in the quality of water. Each parameter has a maximum permissible level which the WHO, ICMR has defined. According to the availability of the data, four features of water have been selected for the proposed research in both the datasets as displayed in Table I [10].

TABLE I. STANDARDS FOR WATER QUALITY PARAMETERS

Parameter	Standard values	Unit weights
pH	6.5-8.5	0.2190
BOD	5 mg/L	0.3723
DO	5 mg/L	0.3723
EC	250 μS/cm	0.3710

1) *pH*: It is one of the important attributes when water quality is considered. According to the WHO standards the permissible pH value ranges from 6.5 to 8.5. In case the pH value goes below 6.5, the water loses the property of making vitamins and minerals in the human organism and if above 8.5, it causes skin irritation and the taste of water becomes salty. The aquatic life cannot survive if the water pH is in the range of 3.5 to 4.5.

2) *Bio-chemical oxygen demand*: BOD is the determination of the oxygen demand for stabilizing industrial and domestic waste. These effluents deposited in the rivers contaminate the water quality which can be determined by BOD. 3mg$L^{-1}$is the maximum allowed limit of BOD. According to WHO, BOD should not exceed 6 mgL^{-1}.

3) *Dissolved Oxygen (DO)*: DO reveal the changes that occur due to the aerobic and anaerobic phenomenon and also gives information on the condition of rivers. The suitable range for DO lies between 5-14.6 mgO_2L^{-1}depending on temperature, altitude and salinity.

4) *Electrical Conductivity (EC)*: It is the measurement of capacity to pass electricity across the water. As the amount of dissolved solids and inorganic chemicals increases the conductivity also increases. So through the conductivity measure, it is decided upon the amount of these chemicals. The maximum permissible value is 250 $\mu S/cm$.

C. Support Vector Machine

In machine learning, there are two types of learning: one is the supervised and the other unsupervised learning. SVM falls in the category of supervised learning model which is associated with analyzing data for classification or regression. Support vector classifier (SVC) is based on SVM, for classification of data into two or more classes. This algorithm was suggested by V. Vapnik and his team. SVC is a discriminating classifier well-defined by a partitioning hyperplane. Given a labeled training data, the algorithm yields the best possible hyperplane which groups new instances. In simple words, given a group of training data, each data is already marked into any of the two classes, the SVM algorithm trains the model in a way a new example arrives it is then categorized into any of the two groups as seen in Fig.1.

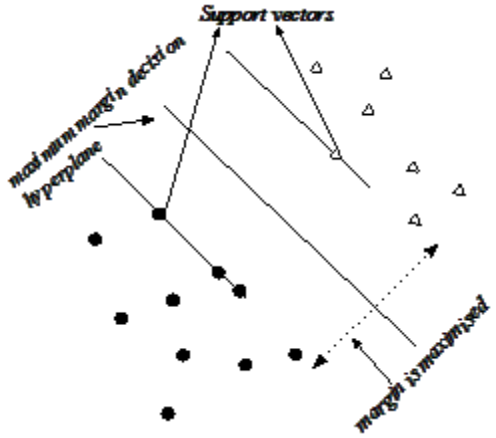

Fig. 1. Support Vector Machine- linearly separable data

Advantages of SVM are:

1) It can be used for the classification of both linear as well as non-linear data.

2) Since it uses a subset of the training data known as support vectors, it is also memory efficient.

3) In the case of classification, the data can be classified into multiple classes depending on the need.

SVC has many kernel functions which are determined according to the classification data. Some main kernel functions are [17]:

I. Linear kernel function: $K(x_i, x_j) = x_i^T x_j$ (1)

II. Polynomial Kernel function: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$ (2)

III. Radial Basis function (RBF):

$$(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0 \quad (3)$$

IV. Sigmoid Kernel: $K(x_i, x_j) = \tanh(\gamma x_i{}^T x_j + \mathrm{r})$ (4)

Where x_i and x_j are the inputs and γ is the regularization factor. The efficiency of the model can be improved by the appropriate selection of kernel and certain parameters such as γ, C and ε.

D. Decision Tree Classifier

Decision Tree Classifier (DTC) is a classification algorithm that keeps on dividing the dataset as in recursive algorithm into smaller sets based on the certain test performed at each node on the tree.

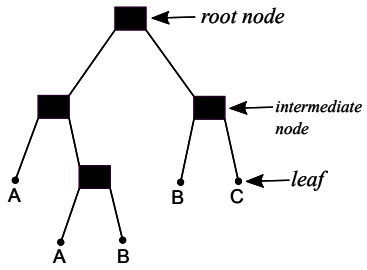

Fig. 2. Decision Tree Classifier

The tree in the DTC consists of a root node (original dataset), intermediate nodes (split datasets) and leaf (final sets of data) as represented in Fig.2. Each tree comprises of one parent node and two or more child nodes. Decision trees have more advantages when compared to traditional classifying methods based on maximum livelihood theory concepts. The decision tree classification is non-parametric and it does not require assumptions on the distribution of input data. Finally, the structure and framework of the decision tree are simple and easy to interpret the classification. Additional advantages are it needs minimal knowledge for data preparation and also achieves good outcomes for large datasets.

Mainly used three algorithms are Iterative Dichotomiser 3, C4.5 and CART algorithm. Out of that the most commonly used is CART (Classification and Regression Tree) algorithm, which was introduced by Breiman et al [11].This algorithm splits a node into two nodes repeatedly based on a predictor variable until the resulting nodes are homogenous enough for the process to terminate. So these homogenous nodes represent the class labels and the intermediate nodes are the features that led to the class labels. Now the splitting criteria correspond to the increase in purity of a node. Three types of splitting criteria are available namely, Gini criterion, Twoing criterion and Ordered Twoing criteria. A few of the stopping rules for splitting the node are if the node becomes pure, if the depth of tree reaches the user-defined tree depth and if the size of the node is larger than the user-defined size[11]. Scikit-learn package in python uses the enhanced version of Breiman's CART algorithm.

E. Naïve Bayes Classifier

Naïve Bayes classification is again an example of supervised learning. It is one of the efficient algorithms. The fundamental assumption of Naïve Bayes is that each of the features is independent and of equal importance. Naïve Bayes is dependent on the Bayes Theorem of probability which is as follows [16]:

$$P(h|X) = \frac{P(h|Y)P(Y)}{P(X)} \quad (5)$$

Where X is the data set, h is the hypothesis such that X falls into a specified class C and $P(h|X)$ is the posterior probability of X [14].

Assume a set of n samples $S = \{S_1, S_2, \ldots\ldots S_n\}$ where the data S_i is of m dimensionality feature vector, $X = \{X_1, X_2, \ldots\ldots X_m\}$. This constitutes the training dataset. Also let the number of classes be k, $c = \{c_1, c_2, \ldots\ldots c_k\}$and every sample belongs to any one of these classes. Introducing an additional data sample X, it can be predicted to which class this particular data sample belongs to using the highest conditional probability, $P(C_i|X)$, i = 1,2…..k.

Using the independence assumption between the features or attributes:

$$P(C_i|X) = \prod_{t=1}^{n} P(X_t|C_i) \tag{6}$$

Where X_t are values for attributes of X. The Naïve Bayes uses some density function such as normal or gauss, lognormal, Poisson and gamma to calculate $P(X_t|C_i)$ values for each attribute. Scikit-learn package in python software uses Gauss distribution and the equation is as follows:

$$P(X_t|C_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{(X_t-\mu_i)^2}{2\sigma_i^2}} \tag{7}$$

Where $-\infty < x, \mu < +\infty, \sigma > 0, \mu$ is mean, σ is the standard deviation.

F. Determination of Water Quality Index

There are mainly four types of water viz. surface water, groundwater, waste water and storm water. Different methods are available to find the water quality index (WQI) for different types of water. One method for storm water is the spatial distribution of WQI. Wastewater quality Index (WWQI) can be found out using different methods such as overall, health and acceptability WWQI. It can be used with different aggregate functions like arithmetic or geometric. Also, WQI for wastewater can be found from arithmetic weighted WQI (WAWQI). National Sanitation Foundation WQI, Canadian council of ministers of the environment WQI and WAWQI are used to find surface and ground WQI. As the study is on surface water, concentration is given to WQI for surface water. Among the three methods specified above WAWQI was found to have better results when compared [12]. WAWQI was proposed by Horton in 1965 and was later established by Brown et al in 1972.The unit weight (W_i) is in inverse relation to the proposed standard values of the corresponding feature.$S_{standard}$ values for each parameter are given in Table I.

$$W_i = K \sum \frac{1}{S_{standard}} \tag{8}$$

The proportionality factor K is determined by:

$$K = \frac{1}{\sum \frac{1}{S_1} + \frac{1}{S_2} + \cdots + \frac{1}{S_n}} \tag{9}$$

Quality rating (Q_i) of each parameter is defined by the following equation:-

$$Q_i = \left(\frac{Q_{actual} - Q_{ideal}}{S_{standard} - Q_{ideal}}\right) 100 \tag{10}$$

Where $Q_{actual}$is the observed value of an ith feature, Q_{ideal} is the ideal value of the feature in pure water, Q_{ideal}= 0 for all with an exception of pH = 7.0 and DO = 14.6 mg/L. Finally, by aggregating all the above equations, the required equation for the WAWQI will be obtained:-

$$WAWQI = \frac{\sum_{i=1}^{i=n} Q_i W_i}{\sum W_i} \tag{11}$$

Where n is the no. of parameters taken into consideration. Once the WAWQI is calculated it has to be classified into different classes based on the range of WQI's received. Table II shows the classification of water depending on the WQI as proposed by Brown et al. [13].

TABLE II. CLASSIFICATION OF WATER BASED ON WAWQI

WQI	Class
0-25	Excellent
26-50	Good
51-75	Fair
76-100	Poor
Greater than 100	Unfit for consumption

III. RESULTS AND DISCUSSIONS

As mentioned, the objective of this work is to compare the performance of the three machine learning models SVM, Decision Tree and Naïve Bayes for water quality classification based on the WQI calculated. Upon implementing the machine learning algorithms, its performance is verified using two datasets.

To compare the efficiency of the three algorithms, two performance evaluation parameters are defined:-

1) **Balanced Accuracy Score:** It is a metric measure to evaluate the performance of the models and is similar to accuracy score. Balanced accuracy is the arithmetic mean of accuracies obtained in each class. It ranges from 0 to 1 with 1 being the highest score. It is mainly used in cases where the classes are imbalanced.

$$Accuracy = \frac{No.of\ correct\ predictions}{Total\ no.of\ predictions\ made} \tag{12}$$

$$Balanced\ Accuracy = \frac{\sum_{i=1}^{n} Accuracy}{n} \tag{13}$$

Where n is defined as the number of classes.

2) **Confusion Matrix:** It is a performance measure for a classification problem in machine learning. The number of correctly classified and incorrectly classified datasets is summarized with proper values. As the name suggests the confusion matrix is a matrix with rows showing the predicted class and the columns showing the actual class of the corresponding data.

As stated in the above section many factors help to optimize the performance of the classifier, these are known as tuning parameters. In the SVM classifier, the two main parameters defining it are the kernel and C parameter. In the training phase, different kernels along with different values of C parameter were tested. Out of the four kernels, linear kernel function was found to be more efficient for this model of data. In the decision tree, the depth of the tree was tested with different values to avoid over fitting.

The results obtained have been depicted in Table III and Table IV. Table III demonstrates the summary of results obtained by using the dataset 1 (Narmada river parameters) with four water quality parameters. From the table, it is seen that out of the three algorithms SVM and DT have better performance when compared to Naïve Bayes .i.e. they have the highest accuracy 87.10% with 132 correctly classified data out of the total of 134 data. The number of misclassified instances is found out from the confusion matrix. Table IV illustrates the results achieved by testing dataset 2 with the above said 4 parameters. From the analysis of the outcomes, it is observed that the Decision Tree algorithm has the highest accuracy of 98.50% among the three algorithms. The number of misclassified instances is low in DT when compared to the instances in SVM and Naïve Bayes.

Comparing both the tables, it is evident that the Naïve Bayes is not an appropriate algorithm for this problem. The number of wrongly classified data is more for the Naïve Bayes model in both datasets. In table III, we see that both SVM and decision tree have the same results whereas in table IV; the count of incorrectly classified data is less in the decision tree model. So by analysis of the results obtained, it can be concluded that the decision tree fits both the datasets well.

TABLE III. CLASSIFICATION RESULTS USING DATASET 1

Classifiers	SVM	Decision Tree	Naïve Bayes
Accuracy	87.10	87.10	74.60
Correctly classified	132	132	129
Incorrectly classified	2	2	5

TABLE IV. CLASSIFICATION RESULTS USING DATASET 2

Classifiers	SVM	Decision Tree	Naïve Bayes
Accuracy	95.63	98.50	95.17
Correctly classified	540	544	531
Incorrectly classified	10	6	19

III. CONCLUSION

This paper presents an analysis of water quality detection using various machine learning models such as SVM, DT and Naïve Bayes based on the weighted arithmetic WQI calculated. The models are tested and verified by using four major water quality factors such as pH, DO, electrical conductivity and BOD of the water. An extensive simulation analysis is performed on two real-time datasets and the results of three machine learning algorithms are presented. Based on the results obtained; Decision Tree algorithm is found to be the most suitable classification model in labeling the quality class of water. The work can be further extended by training the machine learning model with a large dataset and identifying the optimal quality parameters to compute the water quality.

REFERENCES

[1] Forde, Martín, Ricardo Izurieta, and Banu Ôrmeci. "Water and health." Water Quality in the Americas, pp. 27, 2019.

[2] Rahmon, E., "Water for sustainable development", UN Chronicle, vol. 55/1, pp.9-12, 2018.

[3] M. A. Tirabassi, "A statistically based mathematical water quality model for a non-estuarine river system1." JAWRA Journal of the American Water Resources Association, Vol. 7, December 1971, pp. 1221-1237.

[4] Kang, Gaganjot, Jerry Zeyu Gao, and Gang Xie. "Data-Driven water quality analysis and prediction: A survey." IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), 2017, pp. 224-232.

[5] Guo, H. C., L. Liu, and G. H. Huang. "A stochastic water quality forecasting system for the Yiluo River." Journal of Environmental Informatics, vol.1, no. 2, pp.18-32, 2003.

[6] A. Sinha and R. K. Isaac, "An analytical FIS model to check the quality of drinking water," 3rd International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, 2017, pp. 1-6.

[7] W. Xuan, L. Jiake and X. Deti, "A hybrid approach of support vector machine with particle swarm optimization for water quality prediction," 5th International Conference on Computer Science & Education, Hefei, 2010, pp. 1158-1163.

[8] Muhammad, Salisu Yusuf, Mokhairi Makhtar, Azilawati Rozaimee, Azwa Abdul Aziz, and Azrul Amri Jamal. "Classification model for water quality using machine learning techniques." International Journal of software engineering and its applications, vol 9, no. 6, 2015, pp. 45-52.

[9] Haghiabi, A.H., Nasrolahi, A.H. and Parsaie, A., Water quality prediction using machine learning methods. Water Quality Research Journal, vol.53, no.1, pp.3-13, 2018.

[10] Singh, Gurdeep, and Rakesh Kant Kamal. "Application of water quality index for assessment of surface water quality status in Goa." Current World Environment vol. 9, no. 3, pp. 994, 2014.

[11] Breiman L, Friedman J, Stone CJ, Olshen RA, Classification and regression trees, CRC press, 1984.

[12] Gupta, Nidhi, Pankaj Pandey, and Jakir Hussain. "Effect of physicochemical and biological parameters on the quality of river water of Narmada, Madhya Pradesh, India." Water Science, vol 31, no. 1, pp. 11-23, 2017.

[13] Brown, Robert M., Nina I. McClelland, Rolf A. Deininger, and Michael F. O'Connor. "A water quality index—crashing the psychological barrier." In Indicators of environmental quality, pp. 173-182. Springer, Boston, MA, 1972.

[14] Mppcb.nic.in. [online] Available at: <http://www.mppcb.nic.in/proc/narmada-report-2017-18.pdf> [Accessed 17 May 2020].

[15] Anbarivan.N.L, "Indian water quality data," Kaggle, 23-Oct-2018. [Online]. Available: https://www.kaggle.com/anbarivan/indian-water-quality-data. [Accessed: 17-May-2020].

[16] Aiswarya Vijayakumar and A. S. Mahesh, "Quality Assessment of Ground Water on Small Dataset", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 5, 2019.

[17] Aiswarya Vijayakumar and A. S. Mahesh, "Quality Assessment of Ground Water in Pre and Post-Monsoon Using Various Classification Technique", International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2, 2019.