

Water Quality Analysis and Prediction of River Water using Support Vector Machine Model

Prasad M. Pujar¹ and Harish H. Kenchannavar²

¹Assistant Professor Department Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, Karnataka, India

Email: pmpujar@git.edu

²Professor Department Information Science and Engineering, KLS Gogte Institute of Technology, Belagavi, Karnataka, India
Email: harishhk@git.edu

Abstract— Water quality analysis is an important part of the real-time water monitoring system. Ensures water quality is maintained following the Bureau of Indian Standards (BIS). It is a necessary step to control any pollution that could adversely affect the quality of river water and the river ecosystem. Therefore, this paper is an attempt to develop a support vector machine (SVM) model for predicting the water quality of the Krishna River. The analysis of the data is based on the parameters like temperature, potential hydrogen (pH), dissolved oxygen (DO), Nitrate (NO₃), electrical conductivity (EC), total dissolved solids (TDS), biological oxygen demand (BOD), and total dissolved solids (TDS). The proposed model can be considered an effective means to map the status of water quality in the river basin. The analysis is performed by comparing the two kernel functions namely sigmoid, and polynomial, of the SVM technique. The intention here is to measure the efficiency and adequacy of these functions in water monitoring applications. The SVM regression using the ten-fold cross-validation method is applied to the sample data to find the exact drift between the predicted values and the actual value. The root mean square error (RMSE) for the three seasons are analyzed using the different kernel functions and the polynomial function with the efficiency of 97.14 % is found to the suited kernel function for water quality monitoring compared to sigmoid kernel function with the efficiency of 51.43%.

Index Terms— SVM, Kernel functions, Sigmoid, Polynomial, Regression, and RMSE.

I. INTRODUCTION

Water is one of the most important resources concerning social and environmental issues. The issues are mainly due to the urbanization and industrialization process that generates a humongous and wide range of contaminants that mix up with the river water and degrade the quality of the water. The range of contaminants include a.) organic pollutants like hydrocarbons, manure, untreated sewage, etc., b.) heavy metals like lead, arsenic, cadmium, mercury, etc., and c.) bacteriological pathogens. These pollutants can cause dangerous ill effects on all the living beings dependent on them [1]. Thus, the need for real-time water quality monitoring with the capabilities of predicting the water quality by itself depending on the previous analysis is essential. Therefore, the system is equipped with the SVM technique for the prediction. The SVM technique is a machine learning method that specifically makes use of hyperplane for the regression and classification of the data. The data that is near to the hyperplane are called support vectors, basically used for estimating the boundary line in differentiating the two different classes [2].

II. LITERATURE SURVEY

The authors evaluate different machine learning techniques, namely logistic regression (LR), linear discriminant analysis (LDR), support vector machine (SVM), and artificial neural network (ANN) for water quality classification. This experiment was intended to find a suitable classification algorithm for water quality monitoring. According to the analysis carried out, it was found that LR was the most suitable for evaluating water quality, considering the harmonic mean of precision F1 as the best performance measure [3]. In the Water Quality Index (WQI) study, the authors use SVM and two variants of the ANN, namely the feed-forward propagation method (FFBP) and RBF to monitor the wetland over 14 months based on 11 water quality parameters. Based on the analysis, SVM and FFBP were best suited to predict WQI based on R2 values of 99.84% and 99.88%, respectively. Therefore, either SVM or FFBP can be used to predict water quality [4]. An accurate LS-SVM model for WQI has been developed for prediction based on available physical data; the prediction was done with SVM and LS-SVM models. Two independent predictions were made, one with all the parameters and the other with the selected parameters. LS-SVM was found to outperform SVM with 92% accuracy compared to SVM with 91% [5]. The author claims that water quality monitoring around the sugar factories in the East Champaran Bihar region was badly needed due to the human population around the factories. The bore-well is a crucial and basic water supply for the local people for drinking and other anthropogenic purposes. Due to rapid urbanization and industrialization, the groundwater supply is under pressure. The assessment was conducted to better understand the basic quality of the water in the area. Except for the COD parameter, all other water quality parameters were within the standards prescribed by the WHO and the BIS [6]. Water management is considered one of the crucial and important steps that require due attention to control water pollution. Pollution prevention can revitalize the water ecosystem and ensure regulatory compliance. To achieve this, a water quality monitoring system based on SVM has been developed. It is used to predict water quality using six water quality parameters, such as pH, particulate matter (SS), dissolved oxygen (DO), ammonia nitrogen (AN), chemical oxygen demand (COD), and biochemical oxygen demand (BOD). The results of this experiment with the Nu-RBF model were successful predictions with an accuracy of 99.8% [7]. The author suggested that the water in and around the city of Akot in Maharashtra was contaminated with traces of heavy metal ions such as Cd, Cr, Pb, Cu, Ni, and Zn, which are considered the most dangerous for human use. Fourteen sampling sites were selected for water quality assessment, of which only four sampling sites were found to be suitable for daily use. The other sites were contaminated with two or more heavy metal ions and the contamination was much higher than the standards prescribed by the WHO. Therefore, the emphasis on continuous monitoring of water quality is essential [8]. The authors discuss the need to monitor and analyze the physicochemical characteristics of the water body to verify that the water is suitable for human consumption based on WQI. The WQI was calculated using the weighted arithmetic index method by adding the unit weight of each parameter together with a subindices score for each parameter. The analysis was carried out for the summer and winter seasons, the result of the analysis was that the water quality in both seasons was acceptable for daily human activities [9]. An IoT system with decision-making facilities is being developed for water quality monitoring. The experiment showed that the system was much more efficient than the traditional system in terms of reliability and accuracy [10].

III. MATERIAL AND METHOD

The area of study considered is a part of the river Krishna in the southern state of Karnataka. The length of the flow considered is approximately 483 km [11]. The parameters like temperature, pH, DO, EC, BOD, NO3, and TDS were utilized. The water quality prediction model is deliberated based on the machine learning concept of SVM. All three seasons namely summer, rainy and, winter water samples are considered for the prediction.

A. Support Vector Machine SVM

The SVM is a supervised machine learning technique used to perform a.) Classification and b.) Regression. This technique analyzes the data to find a hyperplane that divides the data into two classes. The hyperplane position is determined by the data points called support vectors that maximize the margin of the classes separated by the hyperplane. Removing the support vector point can change the position of the hyperplane, so support vectors are considered the critical elements in the dataset and the technique [1][9].

SVM in Water Quality Monitoring: The main purpose of a support vector machine is to isolate the given evidence in the most ideal way. By the time the isolation is complete, the separation between the closest approaches is known as the edge as shown in fig. 3.1. The methodology consists of choosing a hyperplane with the most extreme imaginable limit between the support vectors in the given information sets [4][7].

The SVM technique can be applied to the water quality monitoring application. Here, the water samples can be separated as a.) Normal and b.) Contaminated according to the water quality parameters included in the study. The classification is made based on RMSE, that is, more the difference between the actual and expected values, that sample is classified as contaminated and a smaller difference is classified as a normal sample. Another factor considered in the classification is the choice of the correct kernel function.

Kernel Functions: SVM has several different kernel functions available like Linear, Polynomial, Radial Basis, and Sigmoid functions. An experiment is carried out to find the most efficient kernel function between a polynomial and sigmoid function for water quality prediction based on the accuracy and misclassification [5].

- a. Polynomial Kernel: Displays the similarity of the vectors when training a dataset in a feature space on polynomials of the original variables used in the kernel. It is given by the following equation:

$$k(x,y) = (c + x^T y)^d \quad (1)$$

where d is the degree of the polynomial.

- b. Sigmoid Kernel: This function is equivalent to a two-layer perceptron neural network model, which is used as a activate function for artificial neurons. It is given by the following equation:

$$k(x,y) = \tanh(\gamma x^T y + c) \quad (2)$$

Where $\gamma = 1/N$, N is the data dimension.

The following figure 3.1 shows the flow chart of the classification process.

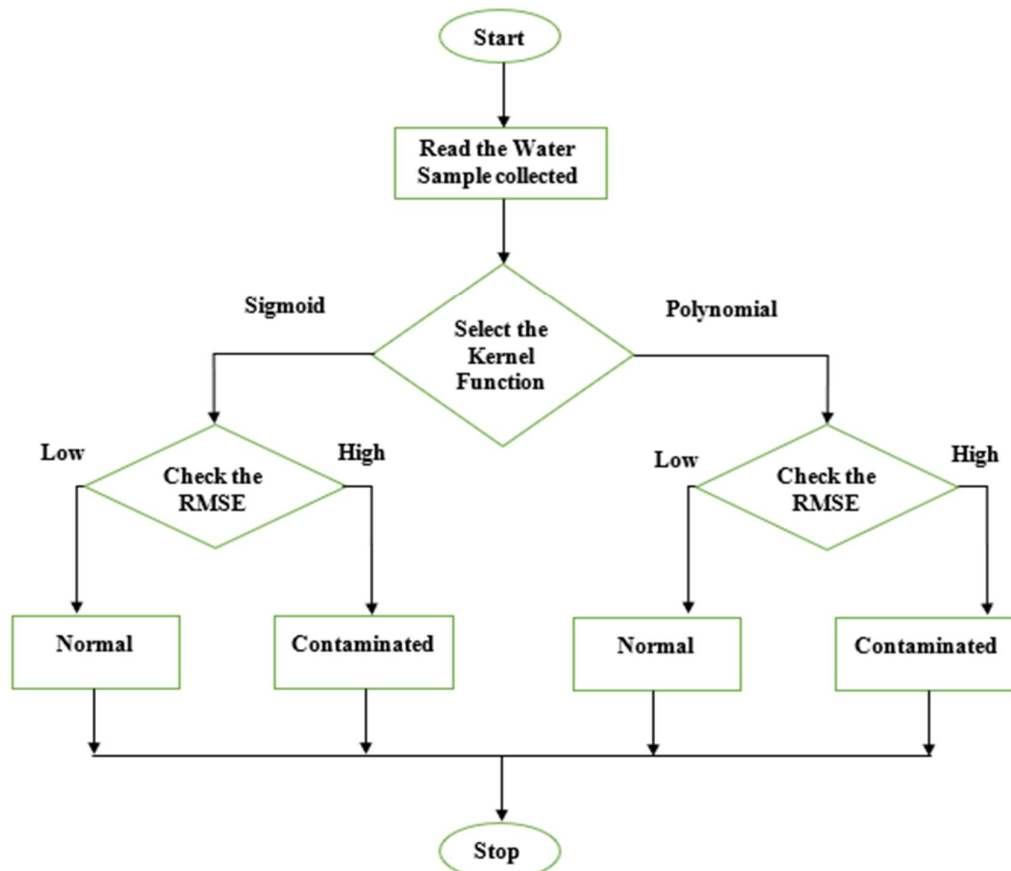


Fig 3.1: Flowchart of the classification process

IV. RESULTS AND DISCUSSION

B. SVM Analysis

From tables 1,2 and, 3 it is observed that for summer season data the efficient classification was observed for polynomial functions with the value of 94.29% whereas the sigmoid function had the worst classification at

42.86%. For the rainy season data, the efficient kernel function was polynomial with a classification rate of 97.14% and the sigmoid function had a classification rate of 40%. And for the winter season, the kernel function polynomial had the classification efficiency of 97.14% as compared to the sigmoid function had a classification done at 51.43%. The result of this experiment shows that amongst the two kernel functions namely polynomial and sigmoid, the polynomial kernel function is more efficient.

The tables show the comparison of these kernel function for the three seasons:

TABLE I: SUMMER SEASON DATA

Type of Kernel	Summer Season	
	Accuracy in %	Mis-Classification in %
Polynomial	94.29	5.71
Sigmoid	42.86	57.14

TABLE II: RAINY SEASON DATA

Type of Kernel	Rainy Season	
	Accuracy in %	Mis-Classification in %
Polynomial	97.14	2.86
Sigmoid	40	60

TABLE III: WINTER SEASON DATA

Type of Kernel	Winter Season	
	Accuracy in %	Mis-Classification in %
Polynomial	97.14	2.86
Sigmoid	51.43	48.57

According to the data available, it is evident that Polynomial kernel functions are suited for water quality monitoring. Further, it was observed that using the SVM model the RMSE for all the input parameters were observed as shown in table IV.

TABLE IV: THE RMSE FOR SVM MODEL

RMSE	Summer	Rainy	Winter
DO	1.14	0.70	0.66
pH	0.33	0.26	0.15
EC	260.18	79.90	118.67
BOD	3.22	0.33	0.43
NO3	6.49	1.17	3.26
TDS	166.52	51.13	75.96

From the above table, it can be observed that for the summer season the RMSE is higher as expected because of the stagnant and low water levels. In the rainy season due to flood-like situations, the RMSE is varying for some parameters i.e., RMSE is low for EC, BOD, NO3 and, TDS but high for DO and pH because of the variations in the levels of water. In the winter season, it is observed that the RMSE values were in between the summer and rainy seasons.

V. CONCLUSION AND FUTURE SCOPE

It is necessary to have a machine learning concept for real-time water quality assessment for in-time alert in case of any anomalies concerning physicochemical parameters. An SVM model is used to detect the accuracy of the prediction of the water quality. A comparison of polynomial and sigmoid kernel functions was made and found that the polynomial kernel function was more accurate with 97.14% in predicting than the sigmoid function with maximum accuracy of 51.43%. In the future two more kernel functions radial basis function (RBF) and linear kernel function can be assessed against the polynomial kernel function to find out the appropriate kernel for water quality monitoring application.

REFERENCE

- [1] R. K. IBRAHIM, H. A. Afan, A. El-shafie, and C. M. Fai, "Water Quality Prediction Model-Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios."

- [2] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, and R. Irfan, "Efficient Water Quality Prediction Using Supervised," *Water*, vol. 11, no. 11, pp. 1–14, 2019.
- [3] Muharemi, F., Logofătu, D., Andersson, C., & Leon, F. (2018). Approaches to Building a Detection Model for Water Quality: A Case Study. *Studies in Computational Intelligence*, 173–183. doi:10.1007/978-3-319-76081-0_15
- [4] Mohammadpour, R., Shaharuddin, S., Chang, C.K. et al. Prediction of water quality index in constructed wetlands using support vector machine. *Environ Sci Pollut Res* 22, 6208–6219(2015). <https://doi.org/10.1007/s11356-014-3806-7>
- [5] Leong, W. C., Bahadori, A., Zhang, J., & Ahmad, Z. (2019). Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). *International Journal of River Basin Management*, 1–8. doi:10.1080/15715124.2019.1628030
- [6] Ranjan R., Water Quality Monitoring of Groundwater Resources around Sugar Factory, Near East-West Champaran Border, Bihar, India, *Res. J. Recent Sci.*, 2(7), 79-81 (2012). spring water is an essential and vital component of our life. it's used for drinking, irrigation, and industrial purposes, etc.
- [7] Abobakr Yahya, A.S.; Ahmed, A.N.; Binti Othman, F.; Ibrahim, R.K.; Afan, H.A.; El-Shafie, A.; Fai, C.M.; Hossain, M.S.; Ehteram, M.; Elshafie, A. Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios. *Water* 2019, 11, 1231.
- [8] Haribhau M.G., Trace Metals Contamination of Surface Water Samples in and Around Akot City in Maharashtra, India, *Res. J. Recent Sci.*, 1(7), 5-9 (2012).
- [9] Tandel, B.N., Macwan, J.E.M. and Soni, C.K. Assessment of Water Quality Index of Small Lake in South Gujarat Region, India. <http://psrcentre.org/images/extraimages/63.%2010.pdf>
- [10] Yuan, Ye and K. Jia. "A Semi-supervised Approach for Water Quality Detection based on IoT Network." *J. Inf. Hiding Multim. Signal Process.* 7 (2016): 858-866.
- [11] P. M. Pujar, H. H. Kenchannavar, R. M. Kulkarni, and U. P. Kulkarni, "Real-time water quality monitoring through Internet of Things and ANOVA-based analysis: a case study on river Krishna," *Appl. Water Sci.*, vol. 10, no. 1, 2020, doi: 10.1007/s13201-019-1111-9.