ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Review on Data Mining Techniques for Prediction of Water Quality

Priya Singh
Department of Computer Science
Guru Nanak Dev University Regional Campus
Jalandhar, India

Pankaj Deep Kaur
Department of Computer Science
Guru Nanak Dev University Regional Campus
Jalandhar, India

Abstract: Data mining is the exploration and scrutiny of large quantity of information that is able to discover meaningful and significant patterns. This paper studies various data mining techniques for prediction of water quality. This paper reviews the models and various evaluation methods that describe and distinguish the classes of water quality. Various data mining techniques like Artificial neural networks, Naïve bayes, Back propogation algorithm, KNN etc has been explored in this paper.

Keywords: Water Quality, Data Mining techniques.

I. INTRODUCTION

Clean, intact, safe and adequate freshwater is essential to all living organism but reducing water quality has become an universal global issue of concern in household use, industrial and agricultural activities. Weather, geological and hydrological natural influences adversely affect the quality of water. . Many industries have polluted the water that affects the water bodies. This results in the reduction of water quality. The water pollution can be defined as one or more harmful substances present in the water to the extent that can cause various problems for living organisms. Water pollution is a means that has damaged various lakes, oceans and river and other water resource. As a result, it is necessary to monitor the quantity and quality of water.

Water quality can be thought of as a physical, chemical and biological characteristics of water which can be used to predict the water quality. Water quality aids to determine the concentration of chemicals present in the water. Various examples have been studied in the field of water quality. In urban areas, water purification technology is used to remove the harmful contaminants from the water before it is distributed to the homes and other activities use.

Water Quality is dependent on the ecosystem as well as human usage such as industrial pollution, sewage, wastewater and more important overuse of water which leads to the lower level of water. Water quality is regulated by measurements done at the original place and by evaluation of water samples from the location, it must be taken into the laboratory for analysis. The main constituents of water quality identification are assortment and evaluation of water samples, the survey and assessment from the systematic outcomes, and describing the coverage of the appropriate site as well as time frame where that sample pattern was seized. It aids to determine spatial and/or temporal deviations in water quality. Researchers have studied various data mining techniques to predict the water quality. Recently, there has been a thriving interest in studying the broad concept of artificial neural networks (ANNs) that imparts an attractive sustitute tool for water quality modeling and forecasting [2]. Artificial Neural Networks (ANN) with Nonlinear Autoregressive (NAR) time series model is used in order to develop a global methodology for adequate water quality prediction and analysis[9].

Many researchers have used smoothing method which helps them to create prediction equation using past collected data by assigning different weights to each datum. The forecast of algae in raw water can provide time period assurance for the activity of contingency brought by enormous movement of algaes which can offer the reassurance of water supply[4].A decision making tree has been used to predict the chlorophyll level. This method is quantitative and showed as "knowledge tree"(the rule which can deal with the forecast factors that affect the change of chlorophyll).

It is important to carry out water quality assessment for analyzing the water quality safety and sustainable development. Fuzzy c-means clustering method and CWQII and many more other methods have been studied and analyzed to access and evaluate the water quality so as to provide the effective measurement to the environment. CWQII helped researchers to find the class of various water parameters. Determination of various parameters to evaluate at which class level it falls under. It determines the fuzzy comprehensive evaluation (FCE) methodology that is based on fuzzy mathematics to evaluate some of the ill-defined, not easy specificable factors[7]. Various techniques are used to classify and predict the water quality that minimizes the time. Data is collected and then extracted from large datasets and classify the quality using machine learning techniques.

The paper is planned as follows. Section II reviews the data mining techniques. Section III discusses the related work. Section IV has made a comparision table of existing data mining techniques, methods and algorithms. Section V contains the gaps in the literature work. Section VI contains the conclusion and future work.

II. DATA MINING TECHNIQUES

Data Mining is the process of turning raw data into appropriate and meningful information. Various researchers have studied and work on data mining techniques to evaluate and classify the water quality.

A. ANN (Artificial Neural Network)

ANN is a classification model which is grouped by interconnected nodes. It can be viewed as a circular node which is represented as an artificial neuron that reveals the output of one neuron to the input of another. The ANN model is helpful in revealing the unexposed interrelationships in the classical information, therefore expiditing the prediction idea, envision and forecasting of water quality. Based on their performance metrics, we use various formulas which have been illustrated in [1].ANN model is definite and systematic enough to make important and relevant decisions regarding data usage.

B. Naïve Bayes

Naïve Bayes is a classification technique which is based on probability theories which entirely demonstrate the characteristics of water quality assessment[6]. Bayes model is easy to use for very large datasets. In other terms, a Naive Bayes assumed that the value of a distinct feature does not related to the presence or absence of any other feature, given in the class variable. It undergoes through following steps:

1. Extract, clean and classify the water quality.
2. Remove large punctuations and split them.
3. Counting Tokens and calculating the probability. This probability is called as posterior probability which is calculated by the formula described in [6].
4. Adding the probabilities and then wrapping up.

C. Decision Tree

Decision tree is one of the predictive modeling technique used in data mining. It aids to divide the larger dataset into smaller dataset indicating a parent-child relationship. Each internal node defined as inner node is labeled with an input feature. The inner nodes which exhibit many types of attribute test, bifurcations exhibit the test outcomes and leaf nodes particularly exhibit the category of a specific type[4]. Decision tree can handle both numerical and categorical data. It is well suited with large datasets. Higher accuracy in decision tree classification technique depicts that the technique can simulate. It is able to optimize variety of input data such as nominal, numeric and textual. It is a successful supervised learning approach which has the capability of extracting the information from vast amount of data based on decision rules.

D. FNN(Fuzzy Neural Network)

ANN is basically associated with the neurons having the capability of storage and processing for the information. FNN is chosen as a algorithm for data mining introducing the artificial neural networks. It describes the integration of fuzzy Fuzzy logic with the neural network. Fuzzy neural network algorithm which deals with the prediction process is composed of five layers named as: input layer, hidden layer, fuzzification layer, fuzzy reasoning layer and reconciliation fuzzy layer. Zhu has explained the structure of FNN [7]:

- The input layer is the main index that affects the teaching quality as it is the input as well as feedback of fuzzy neural network.
- The fuzzification layer is the activity of computing membership function value from input parameters that belongs to fuzzy set.
- The fuzzy reasoning layer is the fundamental part of fuzzy neural network that is basically employed for simulating the operation of fuzzy relational mapping.
- The reconciliation fuzzy layer is the output layer which means "the distribution of value" that is depicted as by "certainty value ".

The fuzzy rules as well as membership functions which can be stated by using neural network (import) and neural network that is generated to maintain the use as fuzzy inference[3]. Eventually, fuzzy rules and membership functions are usually extricated from the neural network (Export) in association compared to that it would be helpful to describe the actual neural network's central rendering and also the operation.

E. Back Propogation Neural Network(BPNN)

ANN consists of interconnected processing units. Each unit is known as neuron. Each neuron will receive an input from another neuron. Weights are assigned to each neuron. These kinds of weights regulate the nature as well as strength and power of the significance involving the interlocked neurons. The respective signals named as indicators tend to be refined from each and every input and then further processed via a weighted sum to the inputs. The BPNN algorithm criteria looks for the error with the method called as steepest descent. The united weights are modified by simply moving on the way to the negative gradient of the energy function by providing emphasis at each and every iteration for evaluating the network performance. Various performance metrics are used for calculating the network error based on specific formulas. This algorithm follows four major steps:

1. Feed forward computation.
2. Applying back propogation at the output layer.
3. Applying back propogation to the hidden layer.
4. Weights updation.

This algorithm will continue its processing until the value of error function becomes too small.

F. KNN

K-nearest neighbor is an algorithm which is used for regression and classifying the quality problems. It considers various parameters which results in the ease of calculation time and predictive power. It uses a vast amount of classes to calculate the likelihood score. When several KNNs share a class, then the weights of other neighbours to it also added together. Result of such added weights is considered to be the likelihood score. These scores are then sorted in order to find the ranked list. Therefore, KNN is a very simple and effective algorithm.

III. RELATED WORK

Sundarambal Palani et.al [1] proposed ANN models to predict water quality parameters whereas salinity, temperature, dissolved oxygen and Chl-a concentrations using continuous weekly measurements at different locations. It has been observed that the GRNN and Ward Net architecture shows the best performance based on different water quality variables. Depending on their performance,Ward Net is the superior architecture for the temperature and salinity models, but the GRNN is superior for DO and Chl-a models. Wen-Heun Chine et.al [2] proposed ANN model with back propogation algorithm which represents a non-linear relationship to conclude and predict the total nutrient concentration in reservoir in Taiwan. The BPNN accesses the concluded results via a complex structure, but does not able to express the relationships by well-defined precise and explicit functions.

Changjun Zhu et.al [3] proposed fuzzy neural network(FNN) model to evaluate and classify outer water quality in suzhou. The FNN model is reliable and effective and can deal with the problem of solitary elements which reflects the water quality at current stage. Therefore, this methodology is not convenient for the assessment of river water quality.

Jinsuo Lu et.al [4] has established a decision-making tree model that is often used to determine the degree of chlorophyll in natural water in coming day. The idea noticed that the prevailing information of chlorophyll performs a crucial role in influencing the following future approaching days chlorophyll level that usually consists using the principle that the prevailing degree of algae has enormous consequences on the transformation of algae. The results indicated that the accuracy of prediction is likely to reach 80%. Miao Qun et.al [5] proposed a extensive water quality identification index (CWQII) approach to evaluate and classify the water quality of Dagu River in Laixi area of Qingdao, China, making use of one year's auditing data of about three intervals which includes water-deficient period of time, water-common period of time and water-rich period of time. It has compared the category of water quality and hence made a comparision. The result implies that the particular river water quality specifications attain the localized water environment function, that is in association with the spatial and temporal distribution and dissemination of pollutants, from the examination process.

LI Chaunqi et.al [6] proposed a water quality stochastic assessment depending on Bayes to assess regularly per month water quality monitoring data during the conventional segments in Three Gorges Reservoir that varies from the year 2004 to 2006. The whole investigation implies that the actual nutrient salt variables have the tremendous effect to the water quality of the reservoir. ZhenXiang Xing et.al [7] proposed a fuzzy comprehensive evaluation model basically depends on entropy weight method (FCE-EW) that was created to measure the precise state condition of underground water quality. The concluded results are correlated with that of RAGABP and PPC, which indicates that the FCE-EW is a accurate strategy to evaluate the water quality. Salisu Yusaf Muhammad et.al [8] developed an appropriate classification model for analyzing as well as classifying water quality in accordance with the machine learning algorithms. The proposed model is examined and after which the performance is compared with different classification models and algorithms in relation to determine the important attributes which are offered in classifying water quality of Kinta River, Perak Malaysia. The concluded observations demonstrate that the lazy model applying Kstar algorithm is the best algorithm to classify the water quality.

Yafra Khan et.al [9] has developed a water quality forecast model using the support of water quality components applying Artificial Neural Network (ANN) and time-series analysis with ANN-NAR. The performance measures such as Regression, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) indicated the best prediction accuracy results with ANN-NAR time series algorithm. Chadaphim Photphanloet et.al [10] proposed an $\alpha -$ trimmed ARIMA model which is often practiced to calculate the BOD value of the up-coming year making use of assortment of BOD data from the past. The accuracy of BOD prediction attained from the proposed α -trimmed ARIMA model is greater than 70% and the results are better than the smoothing method.

Xiang Yunrong et.al [11] deals with the deep study of a water quality prediction model through application of LS-SVM in Liuxi River in Guangzhou. It presents the SVM-PSO model where PSO algorithm is used to obtain parameters. This algorithm is assumed to be effective and simple showing advantage that it provides better solutions related to quality within a accurate time limit. Hamadoun Bokar et.al [12] shows four classes of water chemical types for both shallow and deep groundwater of Changchun city. A regression analyses method has been used to determine the relationship between different anions and depth of water and with other some nitrates. It presents the correlation relationship and contamination index mapping. Scatter analysis has showed both strong correlation and negative correlation with different anions and cations respectively.

S.Wechmongkhonkon et.al [13] has developed a MLP neural network using the Levenberg-Marquardt algorithm is employed to analyze and distribute the water quality of Dusit district canals of Bangkok,Thailand. MLP results with a very high accuracy with the help of which cost and time is minimized.

IV. COMPARISION TABLE

Reference No.	Authors Name	Paper	Year	Technique	Objective	Advantages	Limitations
[1]	Sundarambal Palani	An ANN application for water quality forecasting.	2008	ANN	To predict and envision the quantitative characteristics of water bodies.	To represent and learn linear and non-linear relationships from the data being modeled.	The size of dataset is small. If huge data is available, the technique would provide more valuable predictions.
[2]	Wen-Huen Chine	ANN for water quality prediction in reservoir.	2009	Back propogation algorithm	To simulate the progressive concentration in reservoir.	It obtain a high non-linear relationship to predict the total phosphorous concentration in reservoir.	BPNN algorithm is not taken into use with other water quality parameters.
[3]	Changjun Zhu	Fuzzy neural network model and its application in water quality evaluation.	2009	FNN	To assess the water quality in area of Suzhou.	The method is effective as it measures the evaluation precision.	The technique is not acceptable for any kind of evaluation of river water quality.
[4]	Jinsuo Lu	Data Mining on forecast raw water quality from Online Monitoring Station based on Decision-Making Tree.	2009	Decision Tree	To predict and forecast the chlorophyll level of coming succeding day.	Forecasting outcomes are excellent.	Several divisions as well as fundamental guidelines in the decision tree is continued remains to be unexplained.
[5]	Mian Qun	Application of comprehensive WQI in water quality assessment of river.	2009	CWQII	To depict the class and evaluation of water quality through qualitatively and quantitatively.	The CWQII is effective in evaluating the class of water quality.	It limited the use of comparing the class with other methods.
[6]	LI Chuanqi	Assessment of water quality near the dam area of three Gorges Reservoir based on Bayes.	2009	Assessment Model for Bayes Water Quality	To access the monthly water quality monitoring.	It is simple and effective method and thus have low computational complexity.	Eutrophication problem is still a research area in the upcoming of reservoir management.

[7]	ZhenXiang Xing	Water Quality Evaluation by the Fuzzy Comprehensive Evaluation based on EW Method.	2011	Entropy Weight Method(FCE-EW)	To describe the weight of index of water quality and assess the actual state of underground water.	FCE-EW is easy to employ as it evaluates the water quality.	Training samples are few.
[8]	Salisu Yusaf Muhammad	Classification Model for water quality using Machine Learning Techniques.	2015	ANN, Naïve Bayes, Kstar and J48.	To classify water quality by comparing the performance based on different models.	Kstar algorithm has the best accuracy to classify water quality.	It limited the use of selection of models to find the most sturdy classification model for water quality.
[9]	Yafra Khan	Predicting and Analyzing water quality using Machine Learning:A Comprehensive Model.	2016	Non-linear Auto-regression (NAR) time series algorithm.	It analyses and forecasts the outlook of the values of water quality prediction to determine the concentration of various parameters of water.	Effective performance metrics is evaluated while compared with the previous work.	It limited the use of user-centric approach for better accuracy prediction.
[10]	Chadaphim Photphanloet	Biochemical Oxygen Demand Prediction for Chaophraya River Using Alpha-Trimmed ARIMA Model	2016	ARIMA model	To predict and envision the BOD value of the upcoming data using BOD data from past.	It can be used with both seasonal and non-seasonal time series data.	It limits the use of monthly and day time series data.

Herein, the comparison table shows the various data mining techniques for examine and evaluation of water quality. Most of the data mining techniques depicts the class of the quality among the different classes to which the water quality belongs.

V. GAPS IN LITERATURE

The majority of the pre-existing techniques has certain restrictions and problems, because it has neglected many of the points some of them are:

1. The use of integration of feature selection techniques can be done to enhance the accuracy rate further for recognition of water quality.
2. The majority of the existing techniques are limited to most of the substantial features of water quality.
3. The integration of feature selection technique and Genetic algorithms have been neglected to upgrade the accuracy rate further for recognition of water quality.

VI. CONCLUSION

This paper presents an evaluation for predicting water quality by applying numerous data mining techniques and methods at many different locations. Many existing evaluation methods are studied. Various algorithms have been reviewed for predicting the water quality and hence made a comparison. As a result of analyses, Artificial neural network is used frequently.

REFERENCES

[1] S. Palani, S. Liong, P. Tkalich, “An ANN application for water quality forecasting”, Marine Pollution Bulletin 56 (2008) 1586–1597.

[2] W. Chine, T. Wang, L. Chen, C. Kou, “Artificial Neural Networks for Water Quality Prediction in a Reservoir”, 2009 Second International Workshop on Computer Science and Engineering.

[3] C. Zhu, Z. Hao, “Fuzzy Neural Network Model and its Application in Water Quality Evaluation”, 2009 International Conference on Environmental Science and Information Application Technology.

[4] J. Lu, T. Huang, “Data Mining on Forecast Raw Water Quality from Online Monitoring Station Based on Decision-making Tree”, 2009 Fifth International Joint Conference on INC, IMS and IDC.

[5] M. Qun, G. Ying, L. Zhiqiang, T. Xiaohui, “Application of Comprehensive Water Quality Identification Index in Water Quality Assessment of River”, Global Congress on Intelligent Systems.

[6] L. Chaunqi, W. Wei, “Assessment of the water quality near the dam area of Three Gorges Reservoir based on Bayes”, The 1st International Conference on Information Science and Engineering (ICISE2009).

[7] Z. Xing, Q. Fu, D. Liu, “Water Quality Evaluation by the Fuzzy Comprehensive Evaluation based on EW Method”, 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).

[8] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. A. Aziz, A. A. 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] Y. Khan, C. S. See, “Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model”.

[10] C. Photphanloet, W.Treeratanajaru, N. Cooharojananone, R. Lipikorn, “Biochemical Oxygen Demand Prediction for Chaophraya River Using Alpha-Trimmed ARIMA Model” 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE).

[11] X. Yunrong and L. Jiang, “Water quality prediction using LS-SVM with particle swarm optimization”, Second international workshop on knowledge discovery and Data mining, IEEE, (2009).

[12] H. Bokar, J. Tang, and N. F. Lin, “Groundwater quality and contamination index mapping in Changchun city, China”, Chinese Geographical Science, Vol. 14, No. 1, pp. 63–70, 2004.

[13] S. Wechmongkhonkon, N.Poomtong, S. Areerachakul, “Application of Artificial Neural Network to Classification Surface Water Quality”.

[14] Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, “Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea.,” Sci. Total Environ., vol. 502, pp. 31–41, Jan. 2015.

[15] S. Maiti and R. K. Tiwari, “A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction,” Environ. Earth Sci., vol. 71, no.7, pp. 3147–3160, 2013.

[16] M.J. Diamantopoulou, V.Z. Antonopoulos and D.M. Papamichail “The Use of a Neural Network Technique for the Prediction of Water Quality Parameters of Axios River in Northern Greece”, Journal 0f Operational Research, Springer-Verlag, Jan 2005, pp. 115-125.