Water Quality Analysis and Prediction using Machine Learning Algorithms

Mr.M.Anbuchezhian,
(Regno : 8153 Research Centre : Sri Paramakalyani College, Alwarkurichi)
Associate Professor, P.G.Department of Chemistry, Sri KGS College , Srivaikuntam – 628 619.
(Affiliated to Manonmaniam Sundaranar University, Abishekapatti,Tirunelveli–12,Tamil Nadu, India)

Dr.R.Venkataraman,
Principal, Sri Paramakalyani College, Alwarkurichi – 627 412,
Tirunelveli (D.T), Tamil Nadu, India.

Mrs.V.Kumuthavalli, M.C.A., M.Phil.,
Associate Professor, Department of Computer Science, Sri Parasakthi College for Women, Courtallam – 627 802.
(An Autonomous College of Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli–12, Tamil Nadu, India)

Abstract :

The main objective of this work is to measure water quality using machine learning algorithms. A Water Quality Index (WQI) is a numeric expression used to evaluate the quality of a given water body. In this paper the following water quality parameters were used to evaluate the overall water quality in terms of the WQI. These parameters were as temp, dissolved oxygen (DO) (% sat), pH, conductivity, Biochemical oxygen demand (BOD), nitrates (NO3), faecal and total coli forms (TC). These parameters are used as feature vector to represent the water quality. In paper five kinds of classification algorithms, namely Navie Bayes(NB), Decision Tree(DT), K-Nearest Neighbor (KNN), Support Vector Machine(SVM), and Random Forest(RF) were employed to predict the water quality class. Experiments were conducted using the real dataset containing the details from various places in Tamil Nadu as well as the synthetic dataset generated on the basis of parameters randomly. Based on the performance of five kinds of classifier, it was found out that the Random Forest classifier achieves some improved result compared to other classifiers. From the analysis it shows that machine learning techniques have the good ability to predict the water quality index.

***Index Terms*–Water Quality Index, Water Quality parameters, Data mining, Classification.**

I. INTRODUCTION

The analysis of water quality is compound problem due to the various factors influence in it. In particular, this concept is intrinsically tied to the different intended uses of the water .different uses require different criteria. Lot of research works going on the prediction of water quality. Normally water quality must be defined based on a set of physical and chemical variables that are closely related to the water's intended use. For each variable, acceptable and unacceptable values must then be defined. Water whose variables meet the pre-established standards for a given use is considered suitable for that use. If the water fails to meet these standards, it must be treated before use. Many physical and chemical characteristics can be used to evaluate water quality or the degree of water pollution. Therefore, it is not possible in practice to clearly define water quality either on a spatial or temporal basis by separately examining the behaviour of every individual variable. The alternative, which is also difficult, consists of integrating the values of a set of physical and chemical variables into a unique value (i.e., an overall or global index). A water quality index is defined as a quality index for any use of water by simply determining the specifications required by that use. This indicator included various physical and chemical characteristics. For each variable, the index included a quality value function (generally linear) that expressed the equivalence between the variable and its quality level. These functions were defined using direct measurements of the concentration of a substance or the value of a physical variable obtained through analyses of water samples. The main theme of this paper is to make an analysis of water quality predication using machine learning algorithms with eight kind of parameters such as Ttemperature (Temp), Dissolved Oxygen (DO) (% sat), pH, conductivity, Biochemical oxygen demand (BOD), nitrates (NO3), faecal and total coli forms (TC).

II. RELATED WORK

The water quality index (WQI) is the most commonly used method to classify and communicate existing water quality. In this method, water quality data collected from large and Complex water quality monitoring programs are converted into single numeric values. The WQI values range between 0 and 100, with 100 representing the highest quality. Several studies have demonstrated the efficacy of the WQI at objectively classifying the available water quality data(Cude 2001; Bordalo et al. 2006;

Singh et al. 2008; Cordobaet al. 2010; Ramesh et al. 2010; Lumb et al. 2011;Prasanna et al. 2012). The global adaptability of the WQI lies in its capability of summarizing large amounts of water quality data into simple terms (e.g., excellent, good, bad) and facilitating simple communication to a general audience. In this way, the WQI serves as a benchmark for evaluating management strategies (Akkoyunlu and Akiner 2012).

The formulation of the WQI involves a series of steps that include developing mathematical equations called indices based on observed water quality parameters, assigning a weighting factor to each parameter depending on its importance in the study, and finally applying a suitable averaging formula to arrive at a single numeric value. These steps often make the computations cumbersome and at the same time limit the formulated index to a specific parameters and geographical areas (Cordoba et al. 2010; Mohammad pour et al. 2015). Thus, the application of the WQI to different geographical settings requires modifications regarding the formulations employed, the sets of parameters considered and the overall implementation goals (Abbasi and Abbasi 2012; Golge et al. 2013). Given the complexity of developing a WQI, there is a need to develop an automated system for knowledge extraction from water quality data, which subsequently simplifies the calculation of the WQI and at the same time covers a broad range of water quality criteria for a larger scope of application.

In this paper, an attempt has been made to explore the possibility of the application of predictive techniques of data mining to water quality classification. Different data mining techniques have been used in the literature to emphasize their importance in the environmental domain. Rajagopalan and Lall (1999) applied the k-nearest neighbor (KNN) method to simulate daily precipitation and other weather variables. Bressler et al. (2003) used the decision tree (DT) technique to generate predefined rules for the operation of a reservoir system. Hyvonen et al. (2005) used a wide range of classification methods to identify key parameters needed for atmospheric aerosol particle formation to occur. Palaniet al. (2008) employed an artificial neural network technique to predict and forecast seawater quality. Mucherino et al. (2009)presented a review of k-nearest neighbor, Random forest (RF) and support vector machine (SVM) techniques for variouproblems related to agriculture. Gibert et al. (2010) used knowledge discovery in databases to identify the most characteristicdynamic patterns occurring in a wastewater treatment plant.. Radojevic et al. (2012) identified the factors influencing the number and dynamics of coliform bacteria in natural reservoir waters using a decision tree and cluster analysis. Verma et al. (2013) employed various classification techniques in data mining to construct day ahead, time series prediction models for total suspended solids in wastewater. Kovcs et al. (2014) presented an interesting study that combined the use of a clustering technique and discriminant analysis to mine homogeneous groups of water quality samples from the Neusiedler Sea, the westernmost and the largest steppe lake in Europe. Liu and Lu (2014) compared KNN and SVM techniques to predict total nitrogen (TN) and total phosphorous (TP) from a river location polluted by agricultural nonpoint source pollution. Mohammad pour et al. (2015) predicted water quality index in constructed wetland using SVM technique. The study proved the efficacy of these machine learning techniques, Particularly SVM in successfully predicting the WQI with high accuracy.

This study has been designed to fulfill the following objectives:

1. To develop a predictive model using popular data mining classification techniques to forecast river water quality class.
2. To validate the performance of predictive models using different evaluation metrics and identify the best predictive model for the present study.

III. METHODOLOGY

The proposed system is designed to predict the water quality index. It consists of two phases one as training and another one as testing phase. In both sections the following process are carried out. The block diagram of the proposed system is shown in the figure 1.

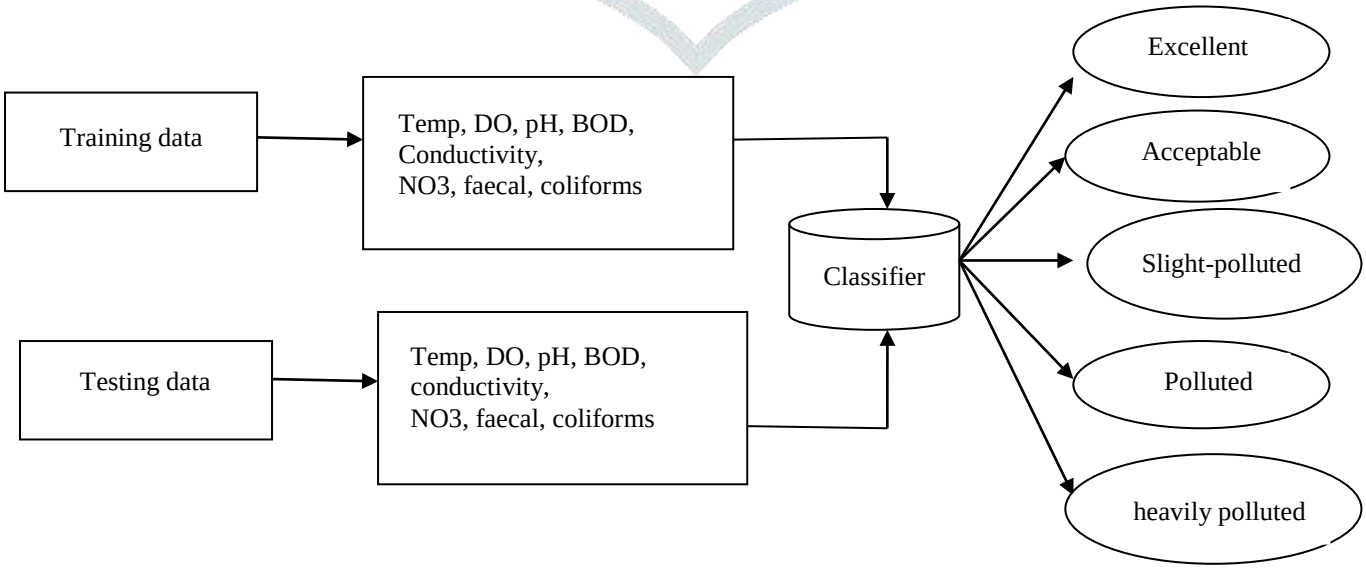

Figure 1 Architecture of the proposed system.

The different steps followed in the execution of the data mining application in this study are as follows:

3.1 Selection of the data set:

Selection of the water quality data set is a prerequisite to model construction and is baon a number of factors such as collection of essential parameters that affect the quality of water, identification of the number of data samples and definition of the class labels for each data sample present in the data. The data sets used in this work consist of 8 indicator parameters. These parameters as temperature, dissolved oxygen (DO) (% sat), pH, conductivity, Biochemical oxygen demand (BOD), nitrates (NO3), faecal and total coli forms (TC). However, the number of parameters and the selection of parameters are not constraints for the proposed approach Corresponding to each data sample in the data set, WQI is first computed and a class label ranging from ‘‘excellent’’ to ‘‘heavily polluted’’ is assigned.

3.2 Designing, learning and testing framework:

The selected data set is used for model learning and evaluation purposes. In this study, a k-fold cross validation technique is used to set the learning and testing framework. Using this technique, the data set is randomly divided into k-disjointed sets of equal size where each part has roughly the same class distribution. Each subset of this division is used in turn as the test set with the remaining subsets being the training set. The performance of the classifier, regarding accuracy, is measured at each step, and all results are averaged to give overall accuracy.

3.3 Building predictive model:

Six different techniques are used to build the predictive model for each training data set created in each iteration of the cross-validation process. These techniques are Naive Bayes(NB), Decision Tree(DT), K-Nearest Neighbour(KNN), Support Vector Machine(SVM), and Random Forest(RF) . All of these techniques have a distinct modus-approach with regard to the underlying relational structure between the indicator parameters and the class label. Hence, it is expected that the performance of each technique will be different for the same data set.

3.4 Evaluating the learned predictive models:

Data mining offers several metrics to validate the performance of different classifiers on some unknown data set. In this study, accuracy, sensitivity, and specificity, precision, recall, f1score are used to evaluate the performances of each classifier.

3.5 Software support

The learning and testing environment was set using a repeated cross-validation technique in the caret package of Matlab software. For implication of the classification algorithm, the following procedure was used:

1. The data set was divided into a training set (85%) and a test set (15%) called D1 and D2, respectively.
2. Repeated cross-validation was applied to the training set with the number of repeats set to 3.
3. Classifiers were trained using the above step.
4. The best parameter setting by the model was identified such that accuracy on D1 was the highest.
5. The model was evaluated on D2, and the performance of the classifiers was recorded using accuracy, sensitivity specificity, precision, recall and f1score as evaluation metrics.

3.6 Classification algorithms

Five data mining algorithms, namely Naive Bayes(NB), Decision Tree(DT), K-Nearest Neighbour(KNN), Support Vector Machine(SVM), and Random Forest(RF) were employed to predict river water quality class. These algorithms belong to a broad category of parametric and nonparametric classifiers, and the purpose of both types of classifiers is to learn a function that maps input variables to output variables from training data set. Since the form of function is unknown, different algorithms make different assumptions about the form of function and the manner in which training data are learnt to produce the output. The classifier following parametric learning style makes stronger assumptions about the data. For these classifiers, if the assumptions come out be correct for any data set, it makes correction decisions. However, same classifier performs badly if the assumptions were wrong.

Common Classifiers that come under this category are: Naive Bayes and rule based. These classifiers do not depend upon size of sample data set in order to learn classification task, rather their working principal are their assumptions. Naive Bayes makes strong assumption that all features present in the data set are independent to each other. Besides parametric nature of this classifier, it is also prone to prediction errors such as bias. Naive Bayes produces high bias, which appears when model makes several assumptions. On the other hand, the rules generated by rule-based classifier have to satisfy mutually exclusive and exhaustive properties. Contrary to parametric learning classifier, nonparametric classifiers do not make any assumptions about the form of the mapping function, and by not making any assumption.

These types of classifiers are free to develop any function form from the training data set (Russell and Norvig 2014). Classifiers considered under this category are SVM, DT, KNN and RF. These classifiers further differ in their approach. SVM, DT and RF are based on learning approaches, whereas KNN works on similarity principle. In other words, SVM, DT and RF classifiers understand the relational structure between features and how group of features influences the outcome variable. More specifically, these classifiers learn about the knowledge that exhibits in the domain that is captured in its given

data set for making future decisions. Large data sets are always an advantage for these classifiers since with the increase in data size, their learning capability also increases. However, small data set provided with complete knowledge on domain is equally beneficial for these classifiers. KNN classifier on the other hand does not learn anything from data rather finds a group of k objects in the training set that is closest to the test object. Unlike SVM, DT this classifier does not rely on knowledge of domain. It simply calculates distance between two features in order to make classification decisions. Since the modus of approach of each selected algorithm is different, evaluation of all these algorithms is important to find out which one is better at approximating the underlying function for same training and testing water quality data sets.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Data generation and collection

Knowledge about the domain is required to make predictions using data mining techniques. For water quality application, it is necessary to knowhow the quality of water is influenced by the various water quality parameters. This knowledge can be gained from the domain expert or from historical data sets. In this work, two types of data sets, a carefully simulated large synthetic data set and an available real data set, were used for the predicting task. The key similarity between the two data sets is that both are tested on equal number of indicator parameters. The data sets are dissimilar on the basis of number of samples considered in each data set. The real data set was limited in terms of number of observations. The synthetic data set was created due to the no availability of large real data sets. However, the designed synthetic data set captures similar relational structures and water quality parameters follow the same distribution as in the real world scenario .For each data set, 8 water quality parameters were used to evaluate the overall water quality in terms of the WQI. These parameters were as temp, dissolved oxygen (DO) (% sat), pH, conductivity, Biochemical oxygen demand (BOD), nitrates (NO3), faecal and total coli forms (TC). The parameter selection was dictated by the fact that they are all commonly monitored crucial parameters, and water quality standards are well defined for these parameters. However, the predictive modelling proposed in this work is flexible enough to work on any number of parameter.

4.2 Synthetic data set

For the purpose of employing data mining algorithms, a target data set is required. As a general practice, if data mining is required as a tool to uncover patterns in the data, then the data set should be large enough to contain these patterns. For a realistic approach to obtain this large data set, a synthetic data set was generated. This synthetic data set was carefully drafted by considering feasible ranges of water quality parameters. The advantage of adopting these concentration ranges was that these ranges have been developed by giving due consideration to water quality standards assigned by various National and International Agencies such as European Union (EU), World health Organization (WHO), Central Pollution Control Board (CPCB) and other reported scientific information's. The index classifies the water quality in five categories namely C1, C2, C3, C4 and C5 where C1 and C5 represent excellent and heavily polluted class, respectively.

The water quality parameters are known to have well defined ranges in which their values can lie; therefore, using these ranges, syntax was developed to randomly generate numerical data for each parameter. The size of the data set was limited to 500 samples under the assumption that this size is large enough to contain the original distributions of indicator parameters. Each sample represented the occurrence of one instance of concentration values of the 8 parameters under consideration. To build a predictive model using classification technique, the data set to be used is required to be supervised in nature. Therefore, the next task was to create a supervised environment for the numerical data set, generated by adding a label to each instance to forecast the pollution level of the water. To achieve this, the WQI was calculated corresponding to each instance of concentration values of the selected 8parameters. The formulation of the Overall Index of Pollution (OIP) from Sargaonkar and Deshpande (2003) was adopted for this purpose.

The OIP is estimated as the average of all of the pollution indices (Pi) for the selected individual water quality parameters and is given by the mathematical expression:

$$OIP = \sum_{i=1}^{n} Pi \qquad (1)$$

where Pi = pollution index of i the parameter, n = number of parameters.

Using the OIP, each instance was labeled as one of five categories, namely excellent (E), acceptable (A), slightly pollute(SP), polluted (P) and heavily polluted (HP). This step prepared the data set for supervised learning. The choice of this particular index was threefold. Firstly, the proposed classification scheme is general and gives due consideration to national and international standards for water quality acceptability under different classes. Secondly, the application of the mathematical formula is simple and did not assign any weight to the water quality parameters, which is often a matter of opinion, thereby making the application of the index subjective (Abbasi and Abbasi 2012). Lastly, the index is validated by a real data set, which is available for citation and can be used to validate the proposed approach also.

Table 4.1 Concentration ranges of water quality parameters

Concentration range	**C1**	**C2**	**C3**	**C4**	**C5**
Class index (score)	1	2	4	8	16
Parameters	**Concentration limits/ranges**				
pH	6.5–7.5	6.0–6.5 and 7.5-8.0	5.0–6.0 and 8.0–9.0	4.5–5.0 and 9.0–9.5	<4.5 and>9.5
DO (% sat)	88–112	75–125	50–150	20–200	<20 and>200
BOD (20 C) (mg/l), max	1.5	3	6	12	24
Nitrates (mg/l), max	20	45	50	100	200
Total coliforms (MPN), max	50	500	5000	10000	150,000

4.3 Real data set

The real datas are collected from various places in Tamilnadu . .For each data set, 8 water quality parameters were used to evaluate the overall water quality in terms of the WQI. These parameters were as temp, dissolved oxygen (DO) (% sat), pH, conductivity, Biochemical oxygen demand (BOD), nitrates (NO3), faecal and total coli forms (TC). The parameter selection was dictated by the fact that they are all commonly monitored crucial parameters, and water quality standards are well defined for these parameters. However, the predictive modeling proposed in this work is flexible enough to work on any number of parameter. Sample data of the real dataset is shown in the table 4.2.

Table 4.2: Water quality parameters

Station Name	**Temp Min-Max**	**DO Min-Max**	**pH Min-Max**	**conductivity Min-max**	**BOD Min-Max**	**Nitrate Min-Max**	**Faecal Min-Max**	**coliform Min-Max**
METTUR TAMIL NADU	29-32	5.7-8.4	7.5-8.8	475-1000	0.8-3	0-0.8	170-2100	390-3200
RN PUDUR	27-33	1.8-7	7.1-8.3	308-684	1.1-5.4	0-0.6	170-2100	400-4600
PALLIPALAYAM	25-33	5.4-7.5	7.3-8.7	330-835	2-6.2	0.2-0.4	210-2200	430-4900

4.4 Results of Performance measures

The performance of the proposed sliding window approach is evaluated using the parameters such as Accuracy, Sensitivity, Specificity, Precision, Recall, and F1score.

True Positives (TP) - These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$\textbf{Accuracy} = \textbf{TP+TN/(TP+FP+FN+TN)} \quad (2)$$

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\textbf{Precision} = \textbf{TP/(TP+FP)} \quad (3)$$

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

$$\textbf{Recall} = \textbf{TP/(TP+FN)} \quad (4)$$

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

$$\textbf{F1 Score} = \textbf{2*(Recall * Precision) / (Recall + Precision)} \quad (5)$$

The following table 4.3 shows the performance of the proposed approach with different classifiers such as Random Forest, Naive Bayes, Decision Tree, K-Nearest Neighbour and Support Vector Machine. Its corresponding graphical representations are show below the following figures figure 2, figure 3 and figure 4.

Table 4.3: Performance measures of Classifiers

METHODS	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	RECALL	F1SCORE
KNN	0.9888	0.972	0.993	0.9736	0.97	0.9723
SVM	0.9976	0.994	0.9985	0.9942	0.994	0.994
DT	0.9976	0.994	0.9985	0.994	0.994	0.994
BAYES	0.9936	0.9841	0.996	0.9844	0.9841	0.984
RF	0.9992	0.998	0.9995	0.998	0.998	0.998

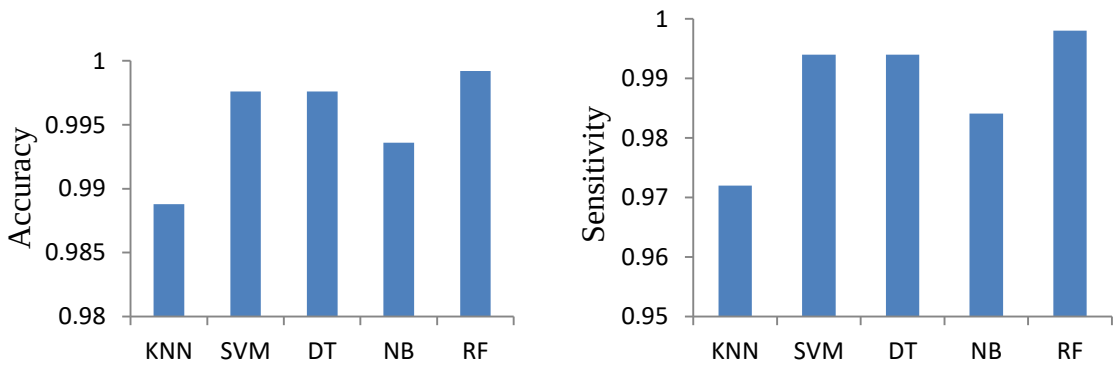

Figure 2 Accuracy and Sensitivity Vs Classifiers

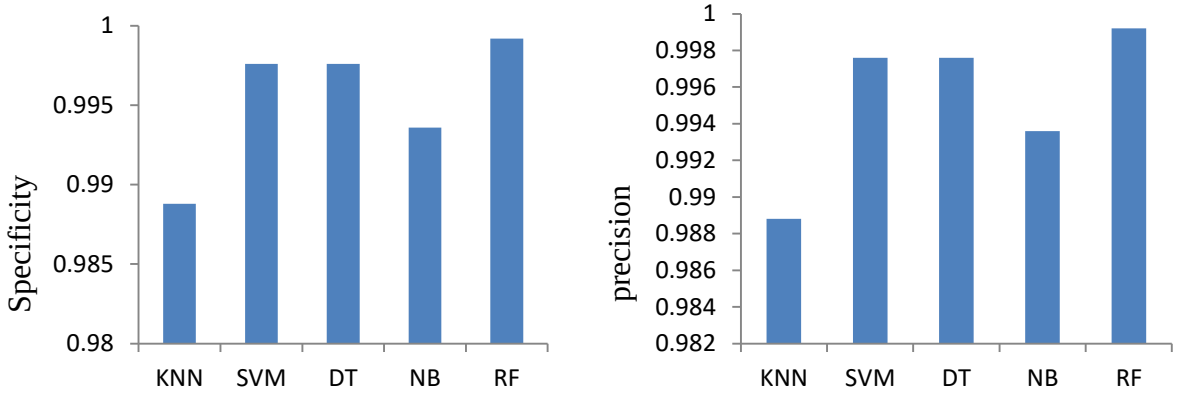

Figure 3 Specificity and Precision Vs Classifiers

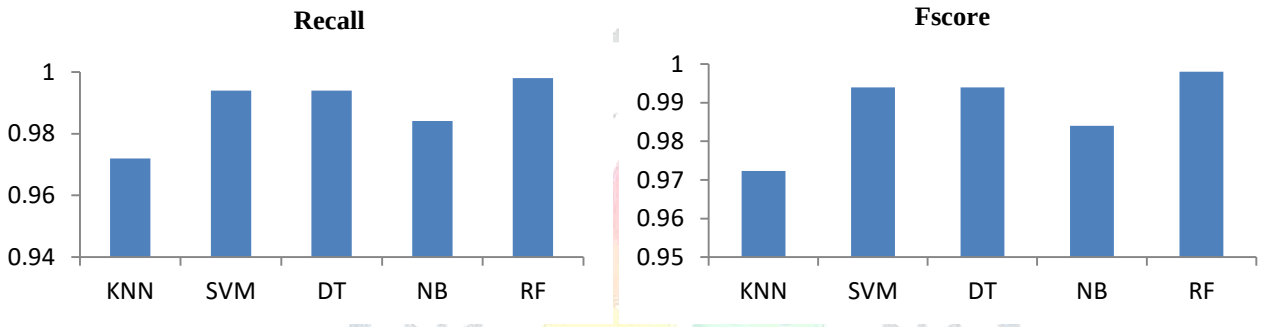

Figure 4 Recall and F1score Vs Classifiers

From the results, it is clear that all kinds of classifier except KNN achieve up to 98% to 99% in terms of all metrics. Among the five kinds of classifier, the Radom Forest classifier attains higher result. DT classifier also obtained the level of RF performance. From this result it is clear that the selected eight type parameters plays vital role in water quality prediction.

V. Conclusion

In this paper the following five well-known data mining classification techniques, namely Naive Bayes, Decision tree, K-nearest neighbour, Support Vector Machines, and Random Forest, were used to classify water quality into excellent, acceptable, slightly polluted, polluted and heavily polluted categories. The models for each classifier had a foundation in the Overall Index of Pollution. The synthetic data set was generated from feasible ranges of 8 water quality parameters: temperature, dissolved oxygen (DO) (% sat), pH, conductivity, Biochemical oxygen demand (BOD), nitrates (NO3), faecal and total coli forms (TC). These ranges complied with both national and international standards. The real data set was obtained from the literature for various places in Tamil Nadu. In the learning phase, the parameters of each classifier were fine-tuned to arrive at the best parameter settings for learning a particular water quality class in the data sets. In the testing phase, each predictive model was validated using unseen data and evaluated by metrics such as accuracy, sensitivity, specificity, precision, recall and F1score. Among the five kinds of classifier Radom Forest classifier attains higher result. DT classifier also obtained the level of RF performance.

References

[1] Abbasi T, Abbasi SA (2012) Water quality indices. Elsevier, Amsterdam.
[2] Akkoyunlu A, Akiner ME (2012) Pollution evaluation in streams using water quality indices: a case study from Turkey's SapancaLake Basin. Ecol Ind 18:501–511. doi:10.1016/j.ecolind.2011.12.018

[3]Bordalo AA, Teixeira R, Wiebe WJ (2006) A water quality index applied to an international shared River Basin: the case of the Douro River. Environ Manag 38:910–920. doi: 10.1007/s00267-004-0037-6
[4]Bressler FT, Savic DA, Walters GA (2003) Water reservoir control with data mining. J Water Res Pl ASCE 129(1):26–34. doi: 10.1061/(ASCE)0733-9496(2003)129:1(26)
[5] Cordoba EB, Martinez AC, Ferrer EV (2010) Water quality indicators: comparison of a probabilistic index and a general quality index. The case of the Confederacion Hidrografica del Jucar (Spain). Ecol Ind 10:1049–1054. doi:10.1016/j.ecolind.2010.01.013
[6] CPCB (2006) Water quality status of Yamuna River 1999–2005: Central Pollution Control Board, Ministry of Environment & Forests, Assessment and Development of River Basin Series:ADSORBS/41/2006-07
[7] Cude CG (2001) Oregon water quality index a tool for evaluating water quality management effectiveness. J Am Water Resour Assoc 37(1):125–137. doi:10.1111/j.17521688.2001.tb05480.x
[8] Gazzaz NM, Yusoff MK, Aris AZ, Juahir H, Ramli MF (2012) Artificial neural network modelling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar Pollut Bull 64(11):2409–2420. doi:10.1016/j.marpolbul.2012.08.005
[9] Gibert K, Rodrguez-Silva G, Rodrguez-Roda I (2010) Knowledge discovery with clustering based on rules by states: a water treatment application. Environ Modell Softw 26(6):712–723. doi:10.1016/j.envsoft.2009.11.004
[10] Golge M, Yenilmez F, Aksoy A (2013) Development of pollution indices for the middle section of the Lower Seyhan Basin (Turkey). Ecol Ind 29:6–17. doi:10.1016/j.ecolind.2012.11.021 Han J, Kamber M (2010) Data mining: concepts and techniques. Elsevier, Atlanta
[11] Hand DJ, Smyth P, Mannila H (2001) Principles of data mining. The MIT Press Cambridge, MA
[12] Hyvonen S, Junninen H, Laakso L, Dal Maso M, Gronholm T, Bonn B, Keronen P, Aalto P, Hiltunen V, Pohja T, Launiainen S, Hari P, Mannila H, Kulmala M (2005) A look at aerosol formation using data mining techniques. Atmos Chem Phys 5:3345–3356
[13] Kovcs J, Kovcs S, Magyar N, Tanos P, Hatvani IG, Anda A (2014) Classification into homogeneous groups using combined cluster and discriminant analysis. Environ Modell & Softw 57:52–59.doi:10.1016/j.envsoft.2014.01.010
[14] Liu M, Lu J (2014) Support vector machine-an alternative to artificial neuron network for water quality forecasting in an agricultural non point source polluted river? Environ Sci Pollut Res 21(18):11036–11053. doi: 10.1007/s11356-014-3046-x
[15] Lumb A, Sharma TC, Jean-Francois Bibeault (2011) A Review of Genesis and Evolution of Water Quality Index (WQI) and Some Future Directions. Water Qual Exp Health 3(1):11–24
[16] Mohammadpour R, Shaharuddin S, Chang CK, Zakaria NA, Ghani AA, Chan NW (2015) Prediction of water quality index in constructed wetlands using support vector machine. Environ Sci Pollut Res 22:6208–6219. doi: 10.1007/s11356-014-3806-
[17] Motamarri S, Boccelli DL (2012) Development of a neural-based forecasting tool to classify recreational water quality usingfecal indicator organisms. Water Res 46(14):4508–4520. doi: 10.1016/ j.watres.2012.05.023
[18] Mucherino A, Papajorgji P, Pardalos PM (2009) A survey of data mining techniques applied to agriculture. Oper Res Int J9 (2):121–140. doi: 10.1007/s12351-009-0054-6
[19] Palani S, Shie-Yui Liong, Tkalich P (2008) An ANN application for water quality forecasting. Mar Pollut Bull 56:1586–1597. doi:10.1016/j.marpolbul.2008.05.021
[20] Prasanna MV, Praveena SM, Chidambaram S, Nagarajan R, Elayaraja A (2012) Evaluation of water quality pollution indices for heavy metal contamination monitoring: a case study from Curtin Lake, Miri City, East Malaysia. Environ Earth Sci 67:1987–2001. doi: 10.1007/s12665-012-1639-6
[21] Radojevic ID, Stefanovic DM, Comic LR, Ostojic AM, Topuzovic MD, Stefanovic ND (2012) Total Coliforms and data mining as a tool in water quality monitoring. Afr J Microbiol Res 6(10):2346–2356. doi:10.5897/AJMR11.1346
[22] Rajagopalan B, Lall U (1999) A k-nearest-neighbor simulator for daily precipitation and other weather variables. Water Resour Res 35(10):3089–3101
[23] Ramesh S, Sukumaran N, Murugesan AG, Rajan MP (2010) An innovative approach of Drinking Water Quality Index-A case study from Southern Tamil Nadu, India. Ecol Ind 10:857–868. doi:10.1016/j.ecolind.2010.01.007
[24] Russell S, Norvig P (2014) Artificial Intelligence: a modern approach. Pearson Education Limited, London
[25] Sargaonkar A, Deshpande V (2003) Development of an Overall Index of Pollution for surface water based on a general classification scheme in Indian Context. Environ Monit and Assess 89:43–67
[26] Singh RP, Nath S, Prasad SC, Nema AK (2008) Selection of suitable aggregation function for estimation of aggregate pollutionindex for River Ganges in India. J Environ Eng-ASCE134(8):689–701. doi:10.1061/(ASCE)0733-9372(2008)134:8(689)
[27] Tan P-N, Steinbach M, Kumar V (2005) Introduction to data mining. Addison-Wesley Longman Publishing Co., Inc, Boston. Verma A, Wei X, Kusiak A (2013) Predicting the total suspended solids in wastewater: a data-mining approach. Eng Appl ArtifIntel 26:1366–1372. doi:10.1016/j.engappai.2012.08.015
[28] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhi-Hua Zhou, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37. doi:10.1007/s10115-007-0114-2