Environ Earth Sci (2017) 76:504
DOI 10.1007/s12665-017-6845-9

ORIGINAL ARTICLE

Predicting river water quality index using data mining techniques

Richa Babbar[1] · Sakshi Babbar[2]

Received: 31 October 2016 / Accepted: 18 July 2017 / Published online: 25 July 2017

Abstract This paper demonstrates the application of data mining techniques to predict river water quality index. The usefulness of these techniques lies in the automated extraction of novel knowledge from the data to improve decision-making. The popular classification techniques, namely *k*-nearest neighbor, decision trees, Naive Bayes, artificial neural networks, rule-based and support vector machines were used to develop the predictive environment to classify water quality into understandable terms based on the Overall Index of Pollution. Experimentation was conducted on two types of data sets: synthetic and real. A repeated *k*-fold cross-validation procedure was followed to design the learning and testing frameworks of the predictive environment. Based on the validation results, it was found that the error rate in defining the true water quality class was 20 and 28%, 11 and 24%, 1 and 38% and 10 and 20% for the *k*-nearest neighbor, Naive Bayes, artificial neural network and rule-based classifiers for synthetic and real data sets, respectively. The decision tree and support vector machines classifiers were found to be the best predictive models with 0% error rates during automated extraction of the water quality class. This study reveals that data mining techniques have the potential to quickly predict water quality class, provided data given are a true representation of the domain knowledge.

✉ Sakshi Babbar
sakshi.babbar@gmail.com

Richa Babbar
richa.babbar@thapar.edu

[1] Department of Civil Engineering, Thapar University, Patiala, Punjab, India

[2] School of Engineering, GD Goenka University, Gurgaon, Haryana, India

Keywords Water quality parameters · Water quality index · Overall Index of Pollution · *k*-fold cross-validation · Data mining classifiers

Introduction

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; Cordoba et 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; Mohammadpour 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. Water quality data are collected at various monitoring stations, and different agencies are responsible for keeping the record resulting in heterogeneity; data sets are often huge that accuracy and comprehensibility becomes accountable; data uncertainty due to imprecise measurement and missing data may be another factor that may contribute to complexity of data use. Under all such circumstances, quick and easy approach to identifying patterns in data sets and classifying data on basis of these patterns can be a useful tool for decision makers. Hence comes the role of data mining techniques from the field of artificial learning. The usefulness of these techniques lies in the automated extraction of hidden predictive information from the data to make knowledge-driven decisions. Several advantages are envisaged by employing a data mining approach to classifying river water quality. Firstly, the approach requires the development of a predictive model, and based on this model, knowledge-driven decisions can be made. Therefore, this obviates any repeated mathematical formulations required to classify water quality from each incoming data sample to draw meaningful conclusions. Secondly, the developed model can be easily modified or re-trained as and when new data samples arrive. The predictive capability of the model will increase with an increase in data sets. Lastly, the model performance can be assessed using various evaluation metrics by which validity of models can be tested for future use.

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. Palani et 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, artificial neural network (ANN) and support vector machine (SVM) techniques for various problems related to agriculture. Gibert et al. (2010) used knowledge discovery in databases to identify the most characteristic dynamic patterns occurring in a wastewater treatment plant. Gazzaz et al. (2012) and Motamarri and Boccelli (2012) used an ANN for water quality classification. 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 ANN and SVM techniques to predict total nitrogen (TN) and total phosphorous (TP) from a river location polluted by agricultural nonpoint source pollution. Mohammadpour et al. (2015) predicted water quality index in constructed wetland using SVM and ANN techniques. 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.

Methodology

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

1. *Selection of the data set*: Selection of the water quality data set is a prerequisite to model construction and is based on 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 10 indicator parameters. These parameters are turbidity, pH, dissolved oxygen (DO) (% sat), biochemical oxygen demand (BOD), total dissolved solids (TDS), hardness, chlorides (Cl), nitrates (NO_3), sulfates (SO_4) and total coliforms (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.

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. *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), DT, KNN, SVM, ANN and rule-based (Rulesb) methods. 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.
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, kappa, sensitivity, and specificity are used to evaluate the performances of each classifier.

Experiment setup

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 know how 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 non-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, 10 water quality parameters were used to evaluate the overall water quality in terms of the WQI. These parameters were turbidity, pH, DO (% sat), BOD, TDS, hardness, chlorides, nitrates, sulfates and total coliforms. 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 parameters.

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 (Hand et al. 2001; Han and Kamber 2010). 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, defined in Table 1, and adopted from Sargaonkar and Deshpande (2003). 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. In order to bring different water quality parameters into commensurate units, an integer value 1, 2 4, 8, 16 is assigned to each class in geometric progression (Abbasi and Abbasi 2012). These numbers are class indices used to categorize the water quality class in numeric value, as shown in Table 1.

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 10 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

Table 1 Concentration ranges of water quality parameters (Sargaonkar and Deshpande 2003)

	Concentration range				
	C1	C2	C3	C4	C5
Class index (score)	1	2	4	8	16
Parameters	Concentration limits/ranges				
Turbidity (NTU)	5	10	100	250	>250
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
TDS (mg/l), max	500	1500	2100	3000	>3000
Hardness as $CaCO_3$ (mg/l), max	75	150	300	500	>500
Chlorides (mg/l), max	150	250	600	800	>800
Nitrates (mg/l), max	20	45	50	100	200
Sulfates (mg/l), max	150	250	400	1000	>1000
Total coliforms (MPN), max	50	500	5000	10,000	150,000

level of the water. To achieve this, the WQI was calculated corresponding to each instance of concentration values of the selected 10 parameters. 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 (P_i) for the selected individual water quality parameters and is given by the mathematical expression:

$$\mathrm{OIP} = \frac{\sum_{i=1}^{n} \mathrm{P}_i}{n} \tag{1}$$

where P_i = pollution index of ith parameter, n = number of parameters.

Using the OIP, each instance was labeled as one of five categories, namely excellent (E), acceptable (A), slightly polluted (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.

Real data set

The real data set used in this study was taken from the Yamuna River in Northern India. The data for this river are available from a previous study (Sargaonkar and Deshpande 2003) for the month of June for the years 1995–1997. The Yamuna River is the largest tributary of the mighty Ganga River and carries the same sanctity as the Ganga River in India. It originates from the Yamunotri glacier at 38°59′N and 78°27′E in the Mussoorie range of the lower Himalayas at an elevation of approximately 6320 m above mean sea level in the *Uttarkashi* district of *Uttaranchal* in Northern India. It travels a distance of 1376 km before its confluence with the mighty Ganga River. Politically, the river travels through seven states of Northern India, and multiple tributaries join the Yamuna River all along its course. The catchment area of the Yamuna River is 366,223 km^2.

The Yamuna River is principally monsoon fed, drawing nearly 80% of its supply during the monsoon period (i.e., July–September). The river water is used for both abstractive and in-stream purposes. Irrigation is the most important use, followed by domestic water supply, industrial and other uses. Several important cities are located on the bank of the river including Delhi, the capital state of India that depends heavily on Yamuna waters for various uses including wastewater discharges. As per a study by the Central Pollution Control Board (CPCB 2006), domestic wastewater is the predominant source of pollution of the river followed by industrial effluents from different categories of industries.

The Central Pollution Control Board has been regularly monitoring the entire stretch of the Yamuna River under the National River Conservation Program (NRCR) and the National Water Quality Monitoring Program (NWQMP). The locations of the monitoring stations used for this study are shown in Fig. 1. The Yamuna waters are of fairly good

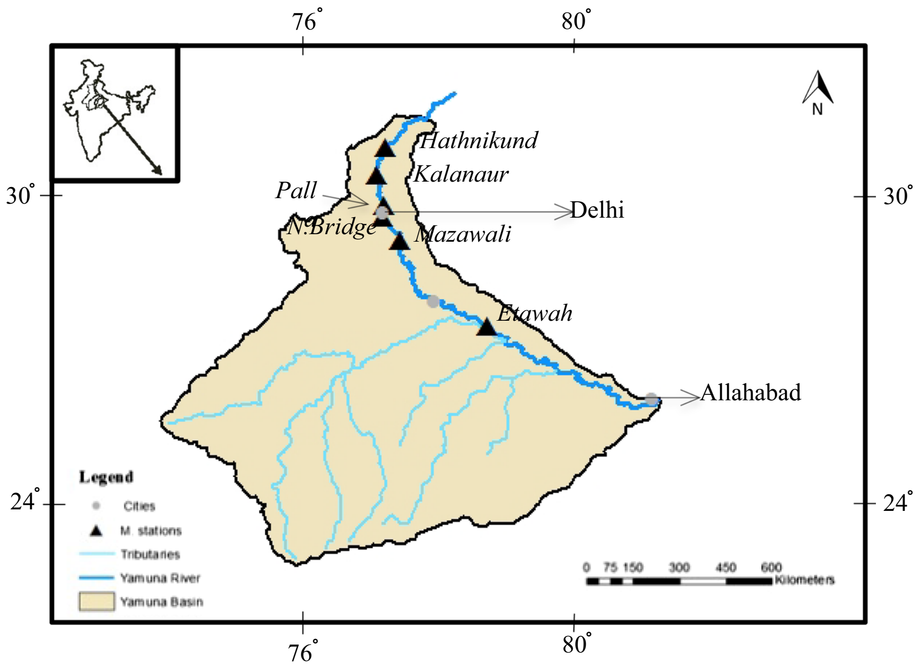

Fig. 1 Yamuna River basin with the monitoring stations used in this study

quality between stations *Hathnikund* and *Kalanaur*. This is because there are no significant wastewater outfalls into the river, and adequate fresh water is available in this river stretch. From *Kalanaur* to the next station, *Palla,* the organic pollution in terms of BOD rises beyond the desired standards. Further downstream, the river stretch is considered worst in terms of pollution level. As many as 17 sewage drains from Delhi empty in this stretch. These drains are located between *Palla* and *Nizamuddin Bridge* stations. The water quality in terms of DO, BOD, and bacteria is not fit for designated best uses in this stretch. In the stretch between *Mazawali* and *Etawah* stations, large withdrawals and discharges of large amounts of untreated and partially treated sewage occur. Beyond *Etawah* station, a significant dilution of pollution load occurs due to the confluence of the clean Chambal River with the Yamuna River.

This study covers the data of these monitoring stations undertaken in the month of June between the years 1995 and 1997. These water quality data are given in Tables 2, 3, 4, 5, 6 and 7.

Software support

The learning and testing environment was set using a repeated cross-validation technique in the caret package of R 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, kappa, sensitivity and specificity as evaluation metrics.

Choice of classification algorithms

Six data mining algorithms, namely NB, DT, KNN, SVM, ANN and Rulesb, 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

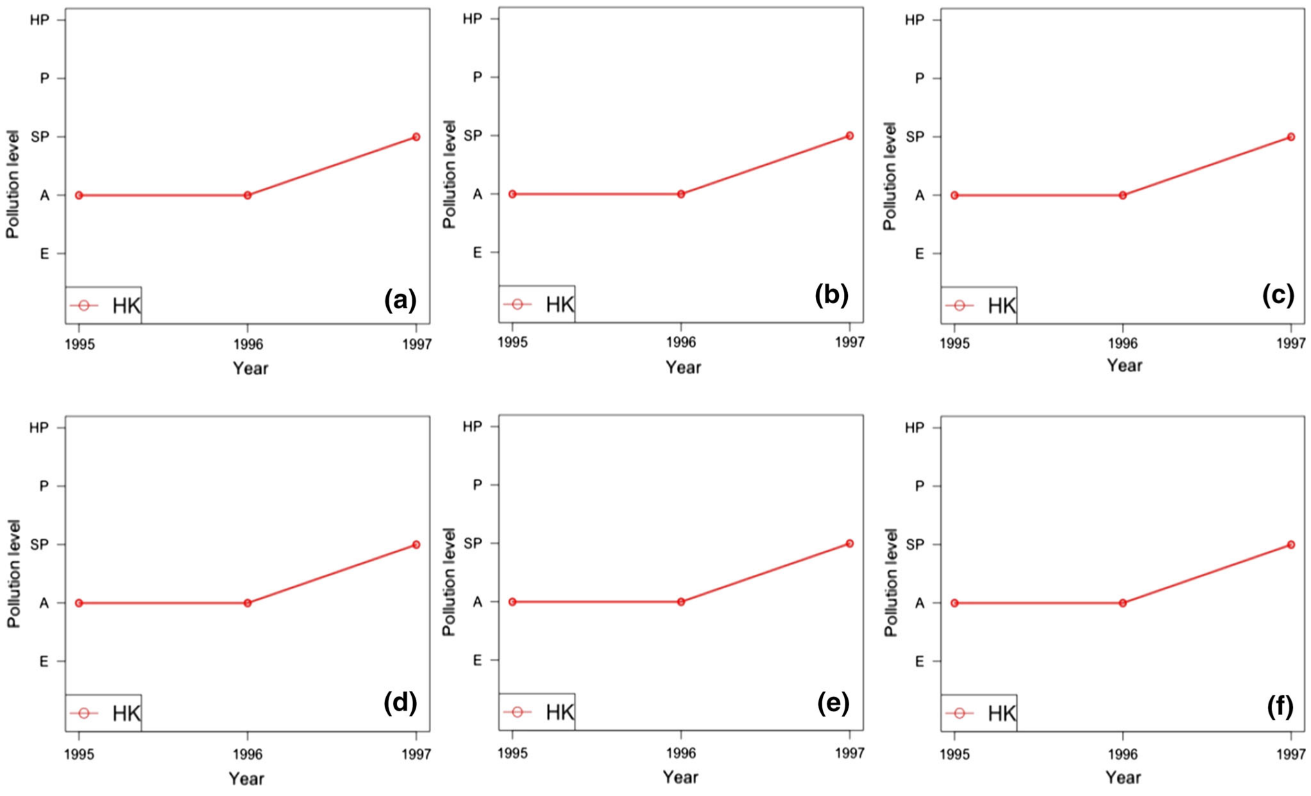

Fig. 2 Water quality class prediction at *Hathnikund* station using **a** KNN, **b** DT, **c** NB, **d** ANN, **e** rule-based and **f** SVM classifiers

Table 2 Real data set for three consecutive years, 1995–1997 at *Hathnikund* (HK) station

Year	Water quality parameters										Water quality class
	Turbidity (mg/l)	pH	DO (% sat)	BOD (mg/l)	TDS (mg/l)	Hardness (mg/l)	Cl (mg/l)	NO_3 (mg/l)	SO_4 (mg/l)	TC (MPN)	
1995	1	7.87	82.02	1	124	91	12	0.002	18	452	A
1996	5	8.22	102.53	1	168	116	9	0.002	42	220	A
1997	3	8.22	123.03	1	212	144	14	0.002	45	3400	SP

Table 3 Real data set for three consecutive years, 1995–1997 at *Kalanaur* (KN) station

Year	Water quality parameters										Water quality class
	Turbidity (mg/l)	pH	DO (% sat)	BOD (mg/l)	TDS (mg/l)	Hardness (mg/l)	Cl (mg/l)	NO_3 (mg/l)	SO_4 (mg/l)	TC (MPN)	
1995	1	8.07	86.86	1	160	136	8	0.009	24	1512	A
1996	9	6.61	86.86	203	154	182	186	0.009	73	104,000	P
1997	5	8.16	117.4	1	148	86	15	0.009	25	3800	SP

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. The mutually exhaustive property says that no more than one rule with different class labels should cover the same instance. Exhaustive property ensures that each observation is at least covered by one rule. The classifier makes necessary assumptions on these properties during learning

Table 4 Real data set for three consecutive years, 1995–1997 at *Palla* (Pl) station

Year	Water quality parameters										Water quality class
	Turbidity	pH	DO (% sat)	BOD (mg/l)	TDS (mg/l)	Hardness (mg/l)	Cl (mg/l)	NO_3 (mg/l)	SO_4 (mg/l)	TC (MPN)	
1995	1	8.5	77.79	1	180	110	10	0.04	15	620	A
1996	6	8.07	116.08	2	199.5	96	10	0.001	36	2700	SP
1997	3.5	8.02	90.95	2	219	150	30	0.001	40	2000	SP

Table 5 Real data set for three consecutive years, 1995–1997 at *Nazimuddin Bridge* (NB) station

Year	Water quality parameters										Water quality class
	Turbidity	pH	DO (% sat)	BOD (mg/l)	TDS (mg/l)	Hardness (mg/l)	Cl (mg/l)	NO_3 (mg/l)	SO_4 (mg/l)	TC (MPN)	
1995	6	7.61	24.17	7	274	116	35	6.43	46	872,000	HP
1996	5	7.13	0.00	9	414.5	180	100	14.7	72	396,000	P
1997	5.5	7.46	48.35	9	555	216	28	5.26	72	534,000	P

Table 6 Real data set for three consecutive years, 1995–1997 at *Mazawali* (MZ) station

Year	Water quality parameters										Water quality class
	Turbidity (mg/l)	pH	DO (% sat)	BOD (mg/l)	TDS (mg/l)	Hardness (mg/l)	Cl (mg/l)	NO_3 (mg/l)	SO_4 (mg/l)	TC (MPN)	
1995	3	7.93	95.07	8	750	284	101	15.24	69	30,000	SP
1996	10	8.06	102.43	8	793	214	96	9.18	50	86,000	SP
1997	6.5	7.72	87.71	7	836	400	35	5.68	64	151,000	P

Table 7 Real data set for three consecutive years, 1995–1997 at *Etawah* (ET) station

Year	Water quality parameters										Water quality class
	Turbidity (mg/l)	pH	DO (% sat)	BOD (mg/l)	TDS (mg/l)	Hardness (mg/l)	Cl (mg/l)	NO_3 (mg/l)	SO_4 (mg/l)	TC (MPN)	
1995	5	8.5	62.74	12	1188	408	322	4.72	98	29,000	P
1996	16	8.69	103.71	6	1008	254	188	0.03	87	2900	SP
1997	10.5	8.65	101.15	3	828	270	213	0.03	75	2500	SP

phase. This classifier is also prone to misclassification when its assumptions on key properties are violated.

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 ANN. These classifiers further differ in their approach. SVM, DT and ANN are based on learning approaches, whereas KNN works on similarity principle. In other words, SVM, DT and ANN 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 and bases the assignment of an outcome on the predominance of a particular class in this neighborhood (Tan et al. 2005; Wu et al. 2008; Mucherino et al. 2009). Thus, this classifier is always dominated by values that have different features present in the data set takes. Unlike SVM, DT and ANN, 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.

Experiment

The experiments consisted of two phases: model learning and model testing. In the model learning phase, the chosen classifiers were evaluated using both the synthetic and real data sets by following the repeated *k*-fold cross-validation procedure. The learning process involved identification of algorithm for a particular classifier, discovering the best parameter settings, optimization of the objective function and minimization of prediction errors of each classifier. Each classifier works on a set of well-defined learning parameters. The aim of objective function was to improve performance of classifier by making low errors during its learning phase by tuning the best value to the respective learning parameters. It took several runs a classifier to stabilize on its performance. In each run, difference between actual outcome and predicted outcome by the classifier was computed to identify quality of the classifier learnt. The iterative process of learning was stopped when performance reached to a level that could not be improved further. The best-tuned values of learning parameters, corresponding to the chosen algorithms, during the model learning phase on synthetic and real data sets are shown in Table 8.

At the model usage or testing phase, each classifier was further subjected to validation by assessing their performance on unseen observations. The performance evaluation of classifiers was calculated using four evaluation metrics: accuracy, sensitivity, specificity and kappa. Accuracy is a standard metric defined as the ratio between accurate predictions and the total number of predictions made. Higher accuracy implies that the classifier makes fewer wrong predictions or misclassifications than correct predictions. However, this measure of evaluation has a limitation in the form of judging the quality of classifiers because it fails to describe how well individual classes were classified. To study performance of classifiers on each class, metrics sensitivity and specificity were used. Sensitivity measures the proportion of positives that are correctly identified, whereas specificity measures the proportion of negatives that are correctly identified. Kappa is a statistical measure that compares an observed accuracy with an expected accuracy. Kappa is always less than or equal to 1. The higher the kappa value, the better the classifier is in terms of accuracy. Based on these performance metrics, each classifier was evaluated first on the synthetic data and then on the real data set.

Table 8 Parameter setting for different classifiers for two types of data set used

Classifier	Algorithm	Parameter	Data set	
			Synthetic	Real
kNN	knn	k	13	5
Decision tree	C5.0	winnow	False	False
Rule based	RIPPER (JRip)	NumOpt	4	2
ANN	nnet	Size and decay	3 and 0.003	3 and 1e^{-04}
SVM	SvmLinear	C regularization parameter	16 (at sigma = 0.08)	1 (at sigma = 0.009)
NB	nb	fl (factor for Laplace correction) and UseKernel	0 and true	0 and true

Table 9 Performance of classifiers for synthetic data set in terms of different metrics

Water quality class	KNN	DT	NB	ANN	Rule based	SVM
E	(0.47, 0.87)	(1, 1)	(0.76, 0.93)	(0, 1.00)	(0.00, 0.97)	(1.00,1.00)
A	(0.00, 0.20)	(1, 1)	(0.87, 0.83)	(1, 0.89)	(0.77, 0.89)	(0.97, 0.97)
SP	(0.00, 0.10)	(1, 1)	(1, 1)	(0, 1.00)	(0,1)	(1,1)
P	(0.96, 0.54)	(1, 1)	(0.99, 0.74)	(1, 0.97)	(0.94, 0.83)	(1.00, 0.97)
HP	(0.55, 0.96)	(1, 1)	(0.73, 0.99)	(1, 1)	(0.85, 0.94)	(0.97,1.00)
Accuracy	0.80	1	0.89	0.99	0.90	0.99
Kappa	0.55	1	0.77	0.97	0.79	0.97

Table 9 presents the performance evaluation of different classifiers used in this study on the synthetic data set. Each entry in the table using notation (*i*, *j*) represents *i* as the sensitivity and *j* as the specificity against each class label present in the data set. The sensitivity and specificity measures help testing performance of the classifier in correctly predicting the actual class and not predicting the class of an item that belongs to a different class. In other words, *i* represent how good a classifier was at detecting the positive class, whereas *j* represents robustness of classifier in terms of no misclassification. Table also shows accuracy and kappa statistics achieved by each classifier.

Consider performance of KNN classifier (Table 9). For class E, the sensitivity of 47% indicates that KNN classifier correctly identified class E 47% of the time. Higher specificity of 87% shows that when it was not expected by the classifier to predict class E it did not do it 87% of time. The results on class E clearly indicates that KNN classifier was not successful in identifying all observations belonging to class E in the data set. The overall results by KNN classifier on all class labels shows poor performance by the classifier. For all classes, either sensitivity is low or specificity is low. Ideally, a classifier should produce high sensitivity and specificity. As observed from Table 9, the KNN classifier has an accuracy of 80% with moderate kappa statistics. The average value of kappa indicates that there is not a very good relationship between accuracy achieved by the classifier and the actual accuracy expected from it. The possible reason for the average performance of the KNN classifier is its approach to predict the class label of a new instance, based on the class most common among its *k*-nearest neighbors. The accuracy and kappa statistics achieved by all six classifiers are also shown in Table 9.

From Table 9, it is observed that the performance evaluation of the rules-based classifier is better and statistically valid with a kappa greater than 70%, which is considered excellent. However, class wise analysis on results of rules-based classifier revealed its poor performance on classes E and SP as its sensitivity is 0%, indicating that classifier was not able to detect any sample of belonging to these classes. However, high specificity for each class shows classifier's low misclassification results.

The kappa statistics for NB and ANN reveal that the accuracies produced by the classifiers are statistically valid. The sensitivity and specificity of these classifiers show that the NB classifier has competence over the ANN when the SP and E classes are considered. Naive Bayes correctly identified class E 76% of time, whereas ANN completely failed in identifying true samples of class E. Similarly, NB achieved 100% success on class SP, while ANN failed in similar condition.

The DT classifier performed extremely well with an accuracy equal to 1 and kappa also equal to 1. This fact statistically strengthens its performance because kappa is 1, which implies perfect agreement between the observed and expected accuracies. The sensitivity and specificity metrics for each class are also highest, indicating correct identification of class label, with no misclassification by the classifier. One of the possible reasons could be that the DT algorithm was able to find mutually exhaustive and exclusive rules for each class to construct a decision tree.

SVM, like DT, performed extremely well in each class with an overall accuracy of 99%. On each class, sensitivity and specificity are nearly 100%. Higher values to these measures show excellent performance of the classifier in correctly predicting the actual classes and not predicting the class of an item that is of different class. The results strengthen the confidence in performance of SVM classifier.

On similar lines, the performance of the different classifiers was also evaluated on the real data set. As observed in Table 10, the classifiers that performed extremely well on the real data set were SVM and DT. The evaluation results of classifiers in Table 10 are further explained by showing their performance on six stations of the Yamuna River. At each station, the detailed analysis is as following. The WQI (All stations have a WQI already computed for the three years monitored (Tables 2, 3, 4, 5, 6, 7) using a step-by-step procedure of finding out the OIP) at *Hathnikund* station shows a degrading trend, with the computed WQI varying from the acceptable, acceptable to slightly polluted class for years 1995, 1996 and 1997, respectively. Figure 2a–f shows the water quality class predicted for *Hathnikund* station by all six classifiers. All models

Table 10 Performance of classifiers for real data set in terms of different metrics

Water quality class	KNN	DT	NB	ANN	Rule based	SVM
A	(1.00, 0.92)	(1, 1)	(0.00, 1.00)	(1.00,1.00)	(1.00, 1.00)	(1,1)
SP	(0.87, 0.60)	(1, 1)	(0.00,1.00)	(0.00,1.00)	(1.00, 0.70)	(1,1)
P	(0.40, 1.00)	(1, 1)	(0.93, 0.49)	(0.60, 0.07)	(0.60,1.00)	(1,1)
HP	(0.00, 1.00)	(1, 1)	(0.58, 0.94)	(1.00, 0.88)	(0.00, 1.00)	(1,1)
Accuracy	0.72	1	0.76	0.62	0.8	1
Kappa	0.56	1	1	0.024	0.73	1

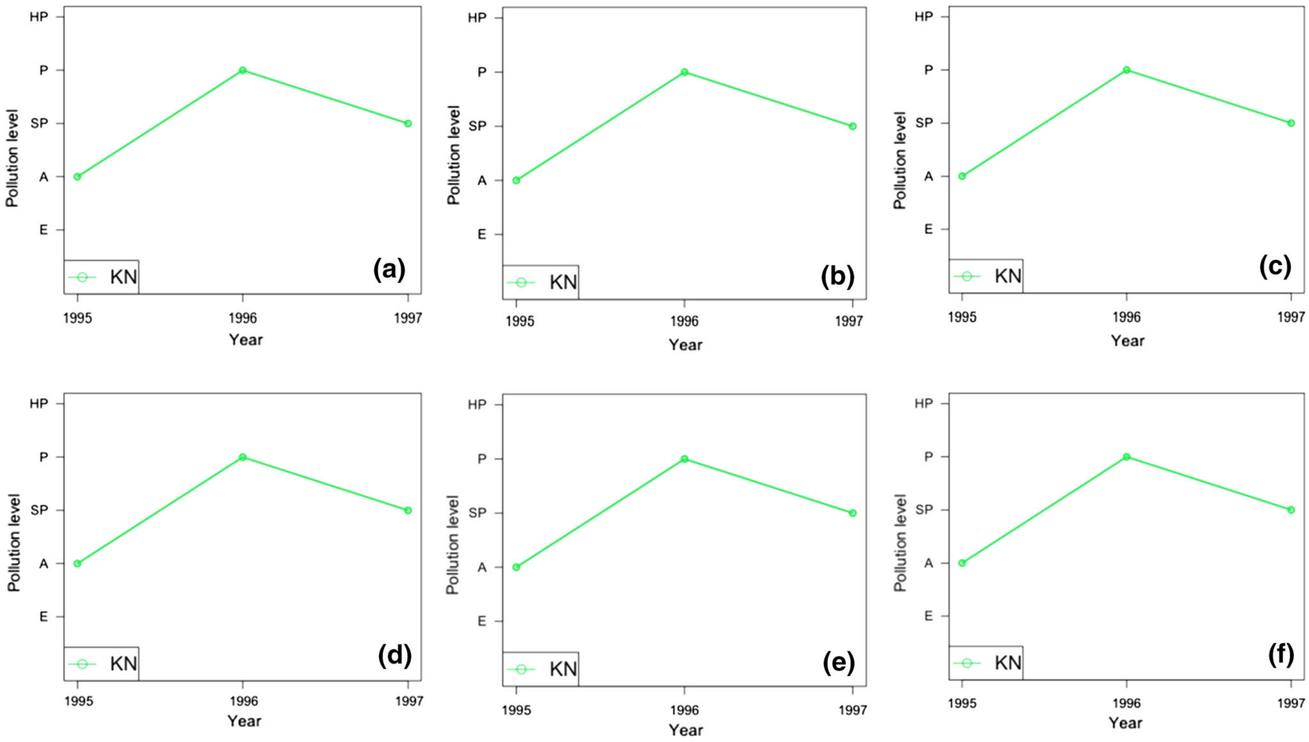

Fig. 3 Water quality class prediction at *Kalanaur* station using **a** KNN, **b** DT, **c** NB, **d** ANN, **e** rule-based and **f** SVM classifiers

classified the real water quality data set of *Hathnikund* station into the acceptable, acceptable and slightly polluted classes for years 1995–1997, respectively. Thus, the predicted results are the same as the observed classes for this station.

The trend at *Kalanaur* station indicates that the water quality class was acceptable in 1995, became polluted in the year 1996 and then improved in 1997. The predicted water quality class profile of *Kalanaur* monitoring station, which is located 36 km downstream of *Hatnikund* station, is shown in Fig. 3a–f. In the case of *Hathnikund* station, all classification techniques were able to accurately predict the water quality class to the classes computed for the three years of data.

At *Palla* station, the water quality class degraded from acceptable to slightly polluted. The computed overall water quality class is observed to be in the acceptable class in 1995 and slightly polluted in 1996–1997 (Table 4). The predicted results of four out of the six classifiers were shown to follow this trend (Fig. 4a–d). The two classifiers KNN and rule-based misclassified the water quality class for the year 1997 as acceptable rather than the slightly polluted class (Fig. 4e, f).

In 1995, water quality at *Nizamuddin Bridge* station was in the heavily polluted class but improved in the following years. Overall, this station was more polluted when compared to other monitoring stations, with the water quality class varying from highly polluted in the year 1995 to the polluted class during 1996–1997 (Table 5). The predicted results are shown in Fig. 5a–f.

Only two classifiers, DT and SVM, were able to correctly predict the true water quality class for all years, while misclassifications were observed in the predicted results for all other classifiers. All classifiers except NB, DT and SVM performed poorly on the HP class. This poor performance is due to a model under-fitting issue. Classifiers under-fit in situations where too few samples are present in a particular class in the data set. In the case of the real data set used in this work, only one sample in the HP class was present; therefore, it was expected that classifiers would perform poorly in this class.

The observed pattern at *Mazawali* station shows continuous degradation from the slightly polluted class to the polluted class. The water quality class at *Mazawali* station was slightly polluted, slightly polluted and polluted for the three consecutive years. Moreover, as is evident from Fig. 6a–f, only the DT and SVM classifiers had 100% accuracy for all three years.

Further downstream at *Etawah* station, all classifiers except KNN and rule-based classifiers provided accurate predictions of the water quality class for all three years (Fig. 7a–f). KNN and rule-based classifiers misclassified the actual polluted class as slightly polluted in 1995 but were able to correctly predict the true class in the subsequent years. The reason for poor performance of rule-based

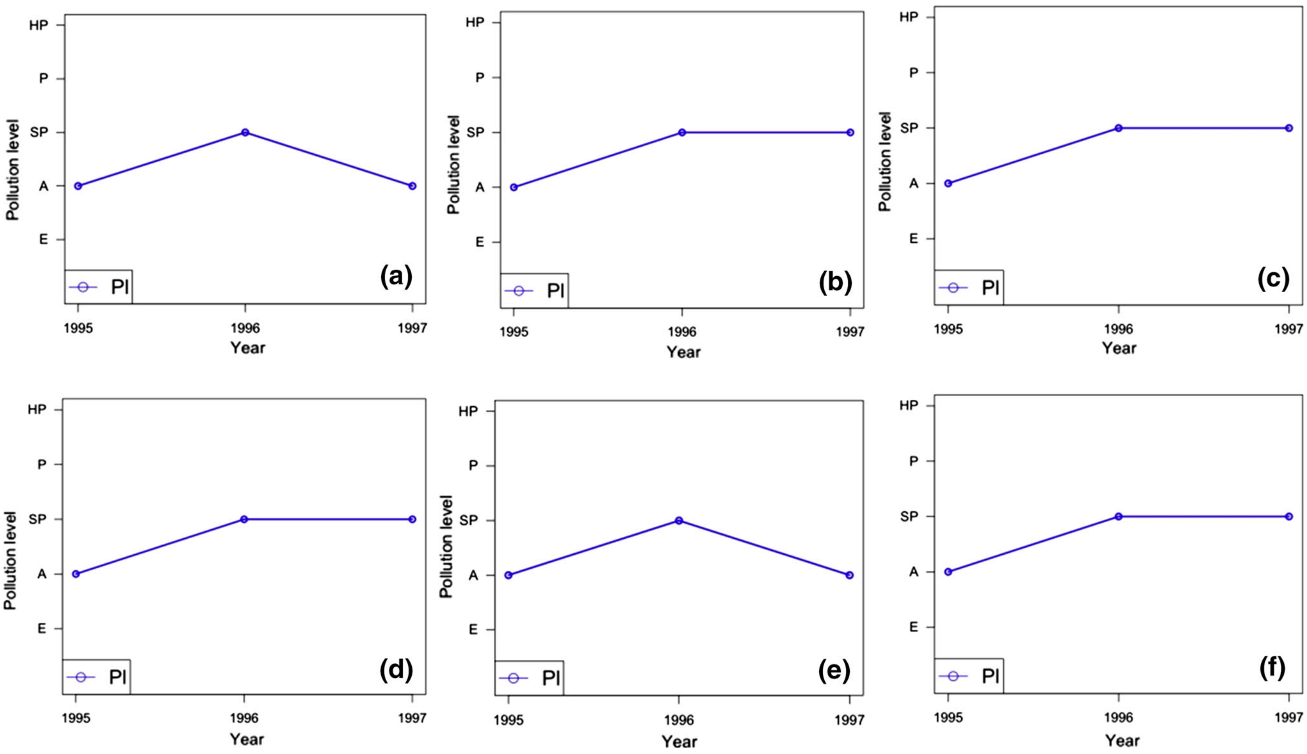

Fig. 4 Water quality class prediction at *Palla* station using **a** KNN, **b** DT, **c** NB, **d** ANN, **e** rule-based and **f** SVM classifiers

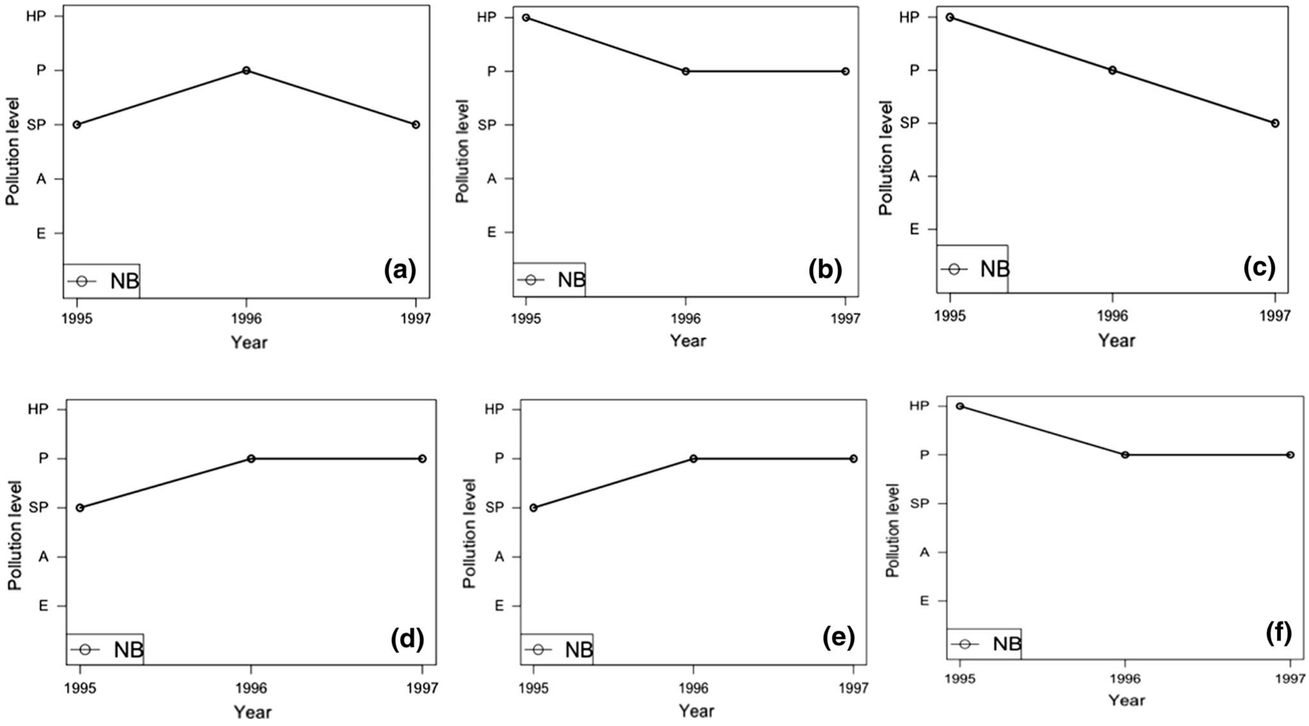

Fig. 5 Water quality class prediction at *Nizammudin Bridge* station **a** KNN, **b** DT, **c** NB, **d** ANN, **e** rule-based and **f** SVM classifiers

classifier is the set of rules extracted from real data set by the rule-based classifier did not cover observation of year 1995 and hence it misclassified. In case of KNN classifier, the reason of misclassification could be that data point was present far away from *k*-nearest neighbor criteria fixed by the classifier.

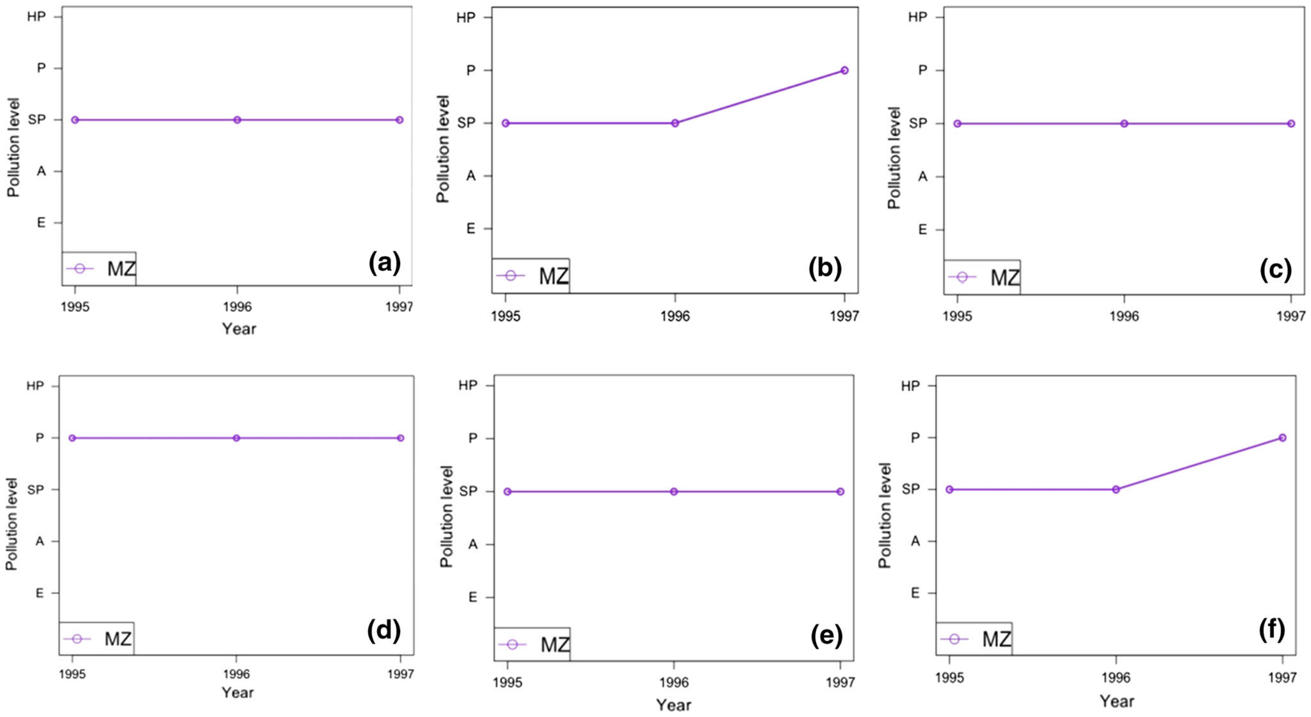

Fig. 6 Water quality class prediction at *Mazawali* station using **a** KNN, **b** DT, **c** NB, **d** ANN, **e** rule-based and **f** SVM classifiers

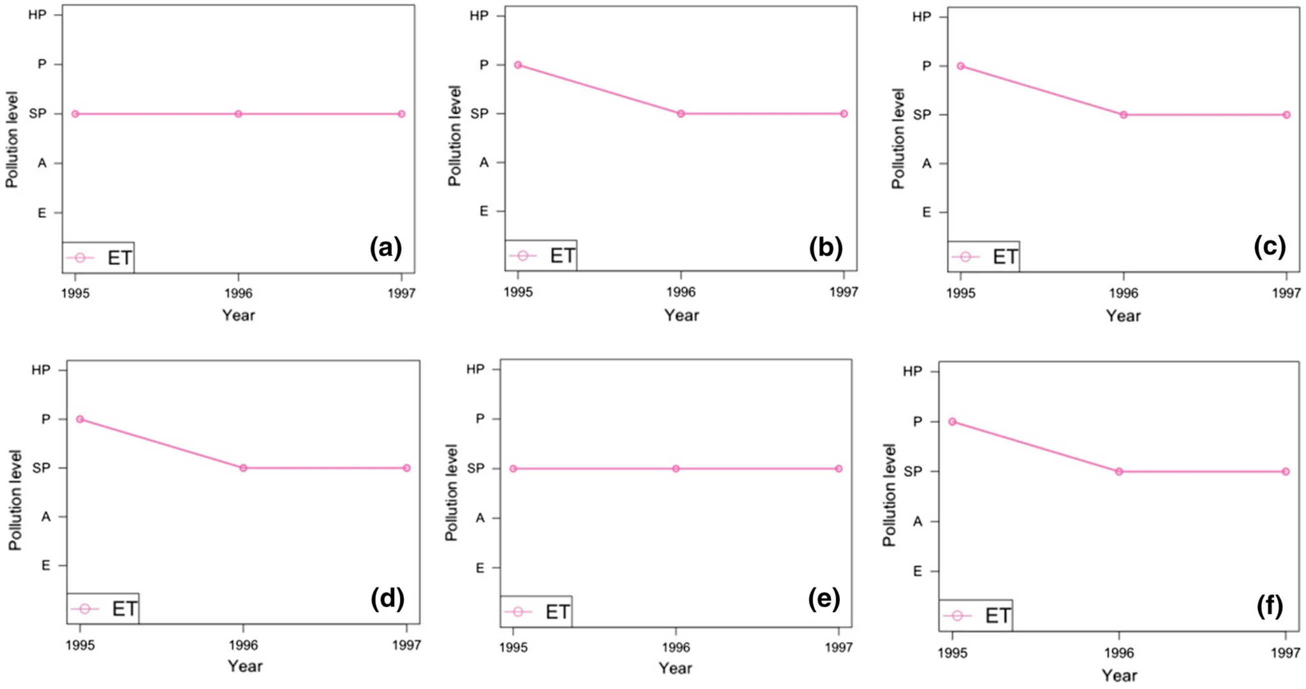

Fig. 7 Water quality class prediction at *Etawah* station using **a** KNN, **b** DT, **c** NB, **d** ANN, **e** rule-based and **f** SVM classifiers

The stability of trained models was evaluated in terms of accuracy and kappa statistics and is shown in Fig. 8. The stability of the performance on the synthetic data set is shown in Fig. 8a. This figure shows that the boxplot for each classifier is very small, indicating that medians of the accuracy and kappa statistics were consistent in every resampling set. It was expected that the SVM, DT and KNN methods would perform well on the test set. On the other hand, small variation in the accuracy and kappa statistics for each classifier is observed from the respective

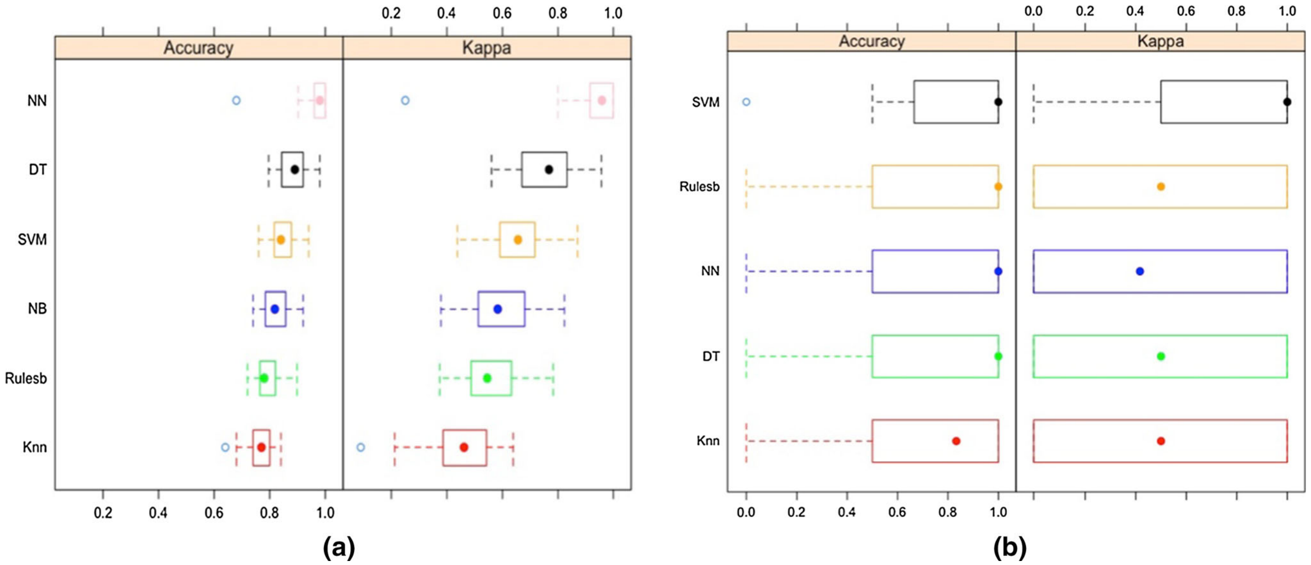

Fig. 8 Stability of classifiers during the *k*-fold cross-validation technique in the **a** synthetic data set and **b** real data set

boxplots in Fig. 8b. Compared to the other classifiers, SVM again showed better stability.

The overall misclassification by each classifier is very low, indicating that all classifiers performed reasonably well. The misclassification is linked to the accuracy achieved in the model testing phase for both the synthetic and real data sets (Tables 9, 10). The misclassification rate or error rate (in percentage) is defined as one minus the accuracy. If calculated for all classifiers, the error rate is found to be greater for the KNN classifier than any other classifier. The error rates for the KNN classifier are 20 and 28% for synthetic and real data sets, respectively. Hence, the predictive capability of the KNN classifier was found to be poorer when compared to SVM and DT, which showed 100% accuracy or 0% error rate on both the synthetic and real data sets. In the case of the NB classifier, the error rates are 11 and 24% on synthetic and real data sets, respectively. The ANN classifier had error rates of 1 and 38%, and the rule-based classifiers had error rates of 10 and 20% on the synthetic and real data sets, respectively.

One very obvious observation from the above discussion is that the results obtained from predictive models based on different classifiers are that their performance is different for two different data sets. This is because the performance of classifiers usually depends on learning style of classifiers, data characteristics and number of samples present in the data set. If parametric and nonparametric styles of learning are considered, then nonparametric learning classifiers namely DT and SVM performed similar on both the data sets. DT and SVM algorithms have the characteristics of learning relationship between features and class labels. Since both data sets captures relationship structure between indicator parameters and water quality index, it was easier for the classifiers to make prediction and remained consistent in performance. ANN being nonparametric behaved different in both the data sets. Unlike SVM and DT, ANN is a slow-learning classifier that requires large data set to learn relational structure present in the data set. In case of synthetic data set containing 500 samples, ANN had 99% of accuracy. Higher kappa statistics authenticate this result, whereas accuracy dropped to 62% for the same classifier on real data set with 18 observations. Very low kappa statistics explains ANN's inability to learn well on a small data set. In case of KNN classifier, kappa statistics is same on both the data sets. However, small variation in accuracy is observed. This indicates classifiers inability in using domain knowledge captured in data to make prediction. As discussed earlier, KNN classifier relies on similarity function to make predictions for unknown observations rather than making predictions based on understanding of the data. Hence, as the number of samples between data sets changed, its results also changed. The probable reason of varied performance of parametric classifiers, namely NB and rule based, on both the data sets can be their underlying assumptions, which may not be completely satisfied to give same performance.

Under the current input settings of the experiment, only two data sets were used on which SVM and DT were identified as the best performers. Since indicator parameters used in this study will always lie in specified ranges given in Table 1, it is assumed that no data set can give more knowledge than one specified in Table 1. While it is possible to drive more data sets from the given knowledge on the parameters, end result would always be the same.

Therefore, whenever a classifier performs good on a data set derived from this knowledge, it will always perform good on any other data set designed or taken considering the knowledge given in Table 1. Thus, by fundamentally establishing the data set on complete knowledge on how water parameters vary on a broad spectrum in environment, it is expected that the results achieved by any data mining technique are reliable and robust.

Conclusions

In this study, six well-known data mining classification techniques, namely Naive Bayes, decision tree, *k*-nearest neighbor, support vector machines, artificial neural networks and rule-based classifiers, 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. This index served as an environment for supervised learning of two types of data sets used in the study. The synthetic data set was generated from feasible ranges of 10 water quality parameters: turbidity, pH, DO (% sat), BOD, TDS, chlorides, nitrates, sulfates and total coliforms. These ranges complied with both national and international standards. The real data set was obtained from the literature for six stations across the Yamuna River in Northern India. The learning and testing environment for predictive modeling was set using a repeated cross-validation technique in the caret package of R software. 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, kappa, sensitivity and specificity. Based on the results, it was concluded that support vector machines and decision tree classifiers proved to be the best classifiers because they achieved statistically valid results. The study also suggested that for data set established on complete knowledge on how water parameters very on a broad spectrum, support vector machines and decision tree classifiers will provide reliable and robust results. This study confirms the potential of data mining techniques in automatic determination of water quality index on a large spectrum of beneficial uses of river water, promulgated by various agencies worldwide.

References

Abbasi T, Abbasi SA (2012) Water quality indices. Elsevier, Amsterdam

Akkoyunlu A, Akiner ME (2012) Pollution evaluation in streams using water quality indices: a case study from Turkey's Sapanca Lake Basin. Ecol Ind 18:501–511. doi:10.1016/j.ecolind.2011.12.018

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

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)

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

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

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.1752-1688.2001.tb05480.x

Gazzaz NM, Yusoff MK, Aris AZ, Juahir H, Ramli MF (2012) Artificial neural network modeling 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

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

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

Hand DJ, Smyth P, Mannila H (2001) Principles of data mining. The MIT Press Cambridge, MA

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

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

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

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

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/s11356-014-3806-7

Motamarri S, Boccelli DL (2012) Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. Water Res 46(14):4508–4520. doi:10.1016/j.watres.2012.05.023

Mucherino A, Papajorgji P, Pardalos PM (2009) A survey of data mining techniques applied to agriculture. Oper Res Int J 9(2):121–140. doi:10.1007/s12351-009-0054-6

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

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

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

Rajagopalan B, Lall U (1999) A *k*-nearest-neighbor simulator for daily precipitation and other weather variables. Water Resour Res 35(10):3089–3101

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

Russell S, Norvig P (2014) Artificial Intelligence: a modern approach. Pearson Education Limited, London

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

Singh RP, Nath S, Prasad SC, Nema AK (2008) Selection of suitable aggregation function for estimation of aggregate pollution index for River Ganges in India. J Environ Eng-ASCE 134(8):689–701. doi:10.1061/(ASCE)0733-9372(2008)134:8(689)

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 Artif Intel 26:1366–1372. doi:10.1016/j.engappai.2012.08.015

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