Modeling Earth Systems and Environment (2018) 4:311–324
https://doi.org/10.1007/s40808-017-0406-9

ORIGINAL ARTICLE

Classification of water quality status based on minimum quality parameters: application of machine learning techniques

Donya Dezfooli[1] · Seyed-Mohammad Hosseini-Moghari[1] · Kumars Ebrahimi[1] · Shahab Araghinejad[1]

Received: 22 August 2017 / Accepted: 12 December 2017 / Published online: 14 December 2017

Abstract

This paper focuses on three models namely probabilistic neural network (PNN), k-nearest neighbor and support vector machine (SVM) as an alternative to NSFWQI in order to classify water quality of Karoon River, Iran as a case study, regarding minimum possible parameters. For this purpose a set of 172 water samples were used in a way that water quality parameters and their water quality classes (from NSFWQI) were considered as the input–output of the models, respectively. Three assessment criteria, namely error rate, error value and accuracy, were applied in order to assess the performance of the applied models. The results revealed that under the condition that no parameter is removed, all the three models showed the same results. However, under quality parameters removal the results revealed that PNN was the best model, as that managed to classify water quality solely based on three quality parameters of turbidity, fecal coliform and total solids, at 90.70% accuracy, 9.30% error rate and error value was 4. In contrast to PNN, in the same condition, SVM showed the poorest performance. As with making use of four quality parameters namely fecal coliform, DO, BOD and total phosphates, it classified water quality at 88.37% accuracy and 11.63% error rate.

Keywords Classification · K-nearest neighbor · Probabilistic neural network · Support vector machine · NSFWQI

Introduction

It is necessary to assess the quality of surface water resources in any region in order to develop agricultural lands, design and operate water resources systems. The ever-increasing growth of the world's population has led to an increase in various forms of water consumption. Not only has it decreased the quantity of available water resources (Rahimi et al. 2017) and imposed more restrictions but it also changed and deteriorated water quality due to urbanization, industrial and agricultural activities (Mohammadi Ghaleni and Ebrahimi 2015; Sharip et al. 2016). Also, Walker et al. (2015) highlighted that it is necessary to assess water quality parameters for better planning and development in terms of water resources management. Various techniques have been proposed in this regard out of which water quality index (WQI) is the most prevalent. Discarding statistic and mathematic complications, WQI is capable of assessing water quality. Regarding this method, the extensive data amassed from measurement of water quality parameters is represented in a single and dimensionless number that has its own defined quality interpretation in accordance with the grading scale (Abbasi and Abbasi 2012; Sadat-Noori et al. 2014). Basically, Horton (1965) introduced the first water quality index. Relying on a simple mathematic method, the index ushered in a whole new era of water quality studies. NSFWQI, OWQI, BCWQI and CCME are among the most prevalent water quality indices. Generally, collection of water samples and laboratory analysis in order to determine the water quality classes of a resource is costly and time-consuming. Therefore, application of Machine Learning Techniques (MLT) could be regarded on the one hand a promising solution to reduce sampling and laboratory costs and on the other hand a possibility to classify water quality in shorter time and by fewer parameters. Sakizadeh (2015) mentioned that aiming at classification machine learning techniques find a mathematic relationship between a set of descriptive variables and a qualitative variable.

Numbers of studies have been conducted regarding application of machine learning techniques in various realms of

✉ Kumars Ebrahimi
EbrahimiK@ut.ac.ir
http://can.ut.ac.ir/member/ebrahimi.aspx

[1] Department of Irrigation and Reclamation Engineering, University of Tehran, Karaj, Iran

hydrology and water resources including streamflow forecasting (Genç and Dağ 2016; Lin et al. 2006; Wu et al. 2009; Aggarwal et al. 2012; Londhe and Panse-Aglave 2015; Parsaie et al. 2015; Kasiviswanathan and Sudheer 2016), rainfall-runoff modeling (Sudheer et al. 2002; Tokar and Markus 2000; Sharma et al. 2015; Javan et al. 2015), predicting groundwater level (Shirmohammadi et al. 2013; Tian et al. 2016; Yoon et al. 2016), drought forecasting (Hosseini-Moghari et al. 2017) and forecasting water level at lakes and dams (Hipni et al. 2013; Lan 2014). Despite the fact that data-driven models have been employed in water quality management and monitoring in recent years (e.g. Hosseini-Moghari et al. 2015; Li et al. 2017; Mohammadpour et al. 2016; Mahmoudi et al. 2016; Sakizadeh 2016), the need for further studies is highly felt. Some researches in this regard are briefly discussed below.

Towler et al. (2009) simulated water quality by applying KNN method. The obtained results revealed that the method was a simple and flexible one capable of classifying various samples at any location. Liu and Lu (2014) applied artificial neural network (ANN) and SVM to forecast total nitrogen (TN) and total phosphate (TP) at any given location in Changle River located in Eastern part of China. Their results indicated that the method is efficient and economical in terms of accurate water quality prediction. Modaresi and Araghinejad (2014) employed three models of PNN, KNN and SVM to classify water quality. Firstly, having applied CCME water quality index, water quality for 100 observed wells were classified into three categories of excellent, marginal and poor based on two parameters of chloride and nitrate. The results revealed that SVM showed the best performance in calibration and validation phases while having the highest error value, KNN showed the poorest performance. Sakizadeh (2015) assessed the performance of linear discriminant analysis and Naive Bayesian classification methods at nine sampling stations along Karaj River, Iran. The results revealed that compared to Naive Bayesian classification method, linear discriminant analysis method delivered a better performance in water quality classification in the studied river. Sakizadeh and Mirzaei (2016) assessed the performance of KNN and SVM in classifying water quality of an aquifer in Khuzestan Province, south Iran. SVM model with 94% accuracy delivered a better performance. Employing of ANN and Gene Expression Programming (GEP), Mohammadpour et al. (2016) made an attempt to forecast the water quality index (WQI) at free surface wetlands. The results revealed that compared to GEP, ANN at 0.988 determination coefficient and 0.013 mean absolute error showed a better performance. They further suggested that these two methods as fast and powerful techniques be used in WQI assessment in order to reduce the required calculations.

The capabilities of data-driven methods in classifying and forecasting water quality have been revealed in previous studies. Given the importance of water quality which plays a significant role in water resources planning, the analysis of water quality classification with fewer water quality parameters, which has not been thoroughly investigated in previous studies, might be of great help to decision-makers dealing with rivers' water quality management. Given the fact that the prevalent water quality indices in surface water quality assessment and monitoring are costly and time-consuming, data-driven modeling methods could be regarded efficient alternatives to overcome these challenges and accelerate decision-makers' reactions in case of troubles or limitations. Having applied three common models of SVM, KNN and PNN, a new approach has been proposed in the present paper, in which the minimum water quality parameters are required to classify water quality. It considerably reduces laboratory and sampling costs and enables experts and researchers to classify water quality in short time and with fewer water quality parameters.

Methods

Case study

Occupying an area of 64,236 km^2, Khuzestan Province is located in the Southeast Iran along with the Persian Gulf and Arvandrood River. Five large rivers of Iran (i.e. Karoon, Karkhe, Dez, Jarahi and Hendijan) are running in Khuzestan province and constitute almost 33 percent of the country's total surface water resource. Substantial amount of industrial and agricultural wastewaters are currently discharging into the above mentioned rivers. It has deteriorated the quality of potable, agricultural and industrial such surface water resources. In the current paper, the qualitative statistics of 172 water samples were used from Gotvand, Arab asad, Shooshtar, Zargan, Ahvaz, Molasani, Koote amir and Darkhovin stations along Karoon River over years 2007–2012 to classify water quality. Table 1 shows their coordinates while their geographical locations are depicted in Fig. 1, below.

Table 1 General characteristics of surveyed stations in Karoon basin

Station	Longitude		Latitude	
	Degree	Minute	Degree	Minute
Gotvand	48	49	32	14
Shooshtar	48	52	31	51
Arab asad	48	51	32	1
Molasani	48	45	31	22
Zargan	48	41	31	20
Ahvaz	48	52	31	35
Koote amir	48	36	31	13
Darkhovin	48	25	30	45

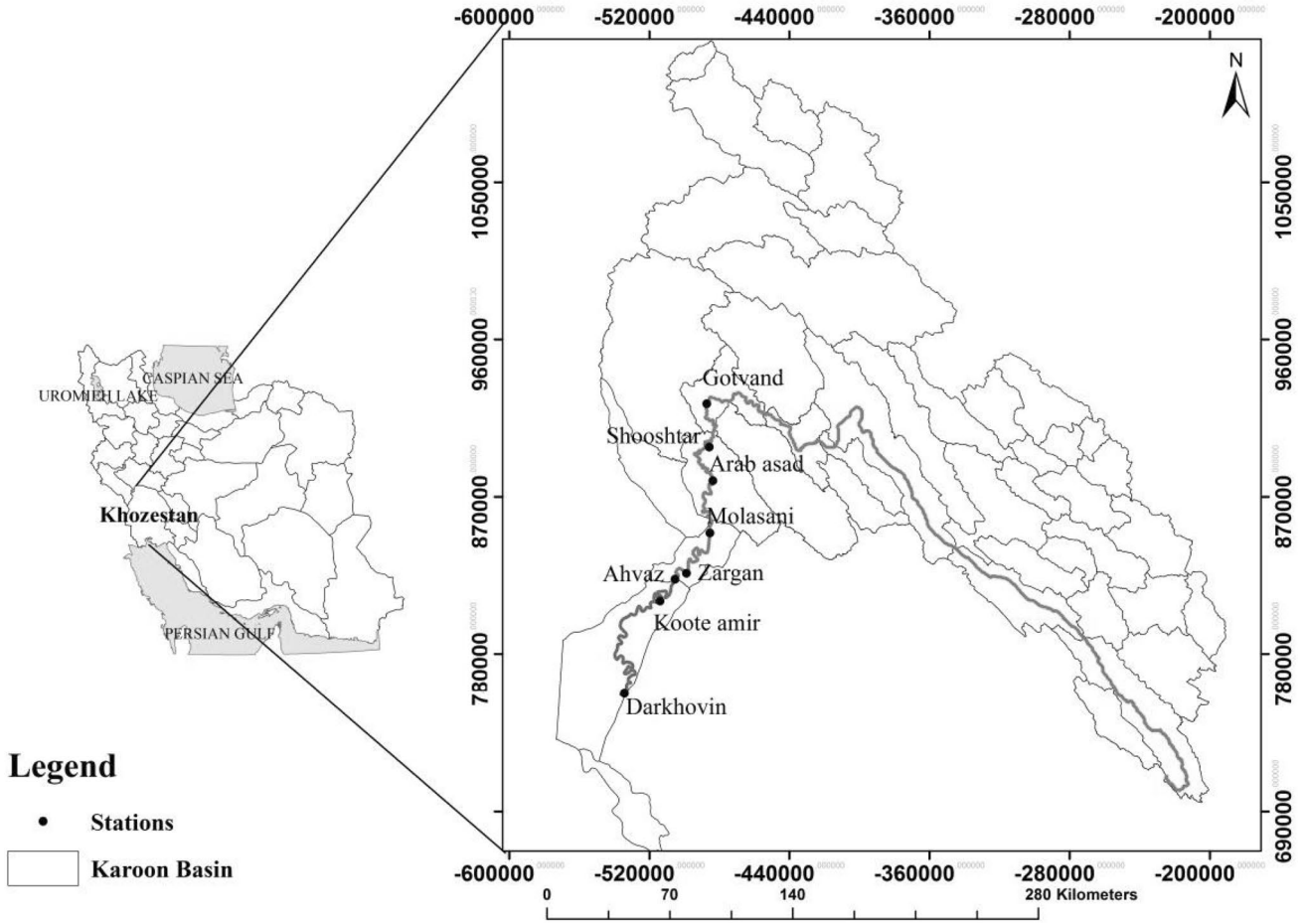

Fig. 1 Location of the study area in Karoon basin, Iran

NSFWQI

Brown et al. (1970) by supporting the National Sanitation Foundation (NSF) proposed a water quality index, involving a survey and getting the ideas of numerous experts, having different expertise in this realm. Firstly, they suggested 35 parameters; subsequent to asking the experts for their opinions, they, however, reduced it to nine, to form the main index. The parameters are included biochemical oxygen demand (BOD), dissolved oxygen (DO), fecal coliform (FC), nitrates, acidity (pH), temperature (Temp), total solids (TS), total phosphates (TP), turbidity (Turb) (for more details see; Landwehr and Deininger 1976; Brown et al. 1972). Having been measured, their sub-indices are obtained from transformation curves. Parameters change to a single number varying from 0 to 100. In this method, each sub-index obtains from an especial relevant curve and then is multiplied by its own weight factor and the total sum, based on Eq. 1, leads to the final index value. Table 2 illustrates the applied weight factors for each parameter in NSFWQI. Finally, the obtained index will be qualitatively classified in accordance with Table 3 (see also Abbasi and Abbasi 2012).

Table 2 Water quality parameters and their weight

Parameter	Weight
Dissolved oxygen (DO)	0.17
Fecal coliform (FC)	0.16
pH	0.11
Biochemical oxygen demand (BOD)	0.11
Temperature	0.1
Total phosphates (TP)	0.1
Nitrates	0.1
Turbidity	0.08
Total solids (TS)	0.07

$$NSFWQI = \sum W_i I_i, \quad (1)$$

where *Ii* is the ith sub-index and *Wi* is the weight factor for ith sub-index.

Table 3 Categorization of NSFWQI

Index value	Category
91–100	Excellent
71–90	Good
51–70	Medium
26–50	Bad
0–25	Very bad

Water quality modeling

Probabilistic neural network (PNN)

Artificial probabilistic neural network (PNN) was firstly introduced by Specht (1990). These networks are used to solve complex problems including approximation, pattern detection and classification. Supervised training is used in this method in order to create distribution functions at the pattern layer. These functions are used to estimate the probability of an input feature vector that is considered a part of a learned category or class (Modaresi and Araghinejad 2014).

PNN classify patterns in accordance with Bayesian strategy and non-parametric probability density function estimators. They are among methods in which statistic patterns existing in data are applied (Masters 1995). Bayesian strategy is applied to a set of strategies and rules in which the expected risk is minimized in order to classify the patterns (Kim et al. 2005). Making use of a set of n-dimensional inputs, a pattern classification technique generally makes decision whether an observation belongs to a specific class. Bayesian decision-making principle to classify a certain pattern (x) and its belonging to a specific class is as follows (Hajmeer and Basheer 2002):

$$x \in c_r \quad if\ h_r l_r f_r(x) \geqslant h_s l_s f_s(x) \quad for\ \forall s, s \neq r\ 1 \leqslant r \leqslant q, \tag{2}$$

where *lr* is the loss related to misclassification of the x pattern to ith class, *hr*is the primary probability of belonging the x pattern to ith class and $f_i(x)$ is the ith class probability density function for the x input.

The structure of the probabilistic neural networks generally includes four layers; an input layer and three information processing one including pattern layer, summation layer and output layer. The input layer's neurons are equal to the number of input factors. No processing is carried out in this layer and it is solely responsible for transferring the input values to all neurons existing in the second layer (Ge et al. 2008). Regarding the pattern layer, the overall number of neurons is equal to the total neurons used to show patterns at each class. Each class may contain lots of training patterns (training vectors) and the number of these patterns is equal to the input factors. The activation function in the pattern layer could be chosen from some kernel density functions; however, Gaussian kernel is often used. The number of neurons is equal to the number of classes in the summation layer. In addition, the activation function is a simple weighted function. Finally, there is only one neuron in the output layer that determines the input class and category (see Modaresi and Araghinejad 2014 and also; Chen et al. 2003). Equation 3 is an example in which Gauss Kernel is applied for each observation of the random variables in order to estimate its density function:

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sigma P}\frac{1}{n_i}\sum_{k=1}^{n_i} e^{\frac{(X-X_{i,k})^T(X-X_{i,k})}{-2\sigma^2}}, \tag{3}$$

where x is the input vector, k is the existing variables in the input vector, n_i is the training patterns existing in the i-th class, $X_{i,k}$ is the k-th training pattern existing in the i-th class and σ is the smoothing parameter (Spread). Making use of Parzen approximation method, the above-mentioned probability density function is directly estimated from training dataset (Cacoullos 1966; Wasserman 1993). Given that PNN is a supervised model, it needs to be trained before being used for classification. At the calibration phase, the smoothing parameter needs to be determined through trial and error in order to reduce misclassifications in training vectors.

K-nearest neighbor (KNN)

KNN could be regarded among most well-known non-parametric models. The K-NN imposes a metric on the predictors to find the set of K past nearest neighbors for the current condition in which the nearest neighbors have the lowest distance (Araghinejad 2013; Jung et al. 2010).

There are various ways to calculate the distance in KNN method; Cosine Distance is, however, used in the present study. If the vector of independent variables is $x_s = (x_1, x_2, \ldots, x_n)$ and the observed past independent variable is $y_t = (y_1, y_2, \ldots y_n)$, the vector difference between current observed independent variables and the past independent variables could be calculated by the following equation (Chomboon et al. 2015):

$$CD = \left(1 - \frac{x_s y_t'}{\sqrt{(x_s x_s')(y_t y_t')}}\right), \tag{4}$$

where CD is cosine distance. Regarding KNN model, the optimum value of parameter K is of paramount importance. In the present study, cross validation was applied in which the best values for parameter K were estimated by an iterative process.

Support vector machine classification

Vapnik (1995) introduced SVM which is a classification model (Araghinejad 2013; Hodge and Austin 2004). Compared to the empirical risk minimization principle used by traditional neural networks, SVM makes use of the structural one which has been proved to be more efficient and can overcome overfitting problems (Liu et al. 2016).

There are many hyperplanes that might classify the data. SVM, however, finds the optimal one that on the one hand minimizes the empirical classification error and on the other hand maximizes the geometric margin (Araghinejad 2013; Li et al. 2014). Suppose that the experimental dataset $\{x_i, y_i\}$, $i=1,\ldots1$ is composed of an instance space $x \in R^n$ and the label set $y=\{-1,1\}$. In the process of training SVM, the following optimization problem solves to find optimal hyperplane.

$$\begin{aligned} &Minimize \frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i \\ &Subject\ to:\ y_i(w^T\phi(x_i)+b \geqslant 1-\xi_i, \quad \forall i \in \{1,...,n\} \\ &\qquad \xi_i \geqslant 0, \quad \forall i \in \{1,...,n\}. \end{aligned} \tag{5}$$

Here,$\phi$is a map from the input space to a feature space, w, $b$and $\xi_i$are the parameters that should be optimized in the training phase. Dual formulation of this function is expressed as follows,

$$\begin{aligned} &Maximize -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i,x_j) + \sum_{j=1}^{n}\alpha_j \\ &Subject\ to:\ \alpha_i \geqslant 0, \quad \forall i \in \{1,\ldots,n\} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \end{aligned} \tag{6}$$

where, $K(x_i,x_j) = (\phi(x_i),\phi(x_j))$ is called the kernel function.

Sigmoid, Radial Basis Function (RBF), Polynomial and Linear are among kernel functions used in the structure of SVM (Cherkassky and Ma 2004; Martins et al. 2016; Barzegar et al. 2016). RBF is used in the present study due to its satisfactory performance. The function could be defined by the following equation:

$$k(x_i,x_j) = \exp(\gamma||x_i - x_j||^2), \tag{7}$$

where γ is the kernel function parameter and its optimum value is calculated through trial and error at the calibration phase.

Subsequent to the specification of optimal parameters α, the decision function of classification for the j-th element becomes,

$$f(x_i) = sign\left(\sum_{i=1}^{n}\alpha_i y_i.k(x_i x_j) + b\right). \tag{8}$$

Multiclass classification with SVMs Multiclass SVM was used in the present study in order to classify the quality of water. There are various ways to apply this method; however, according to the research carried out by Weston and Watkins (1998), their accuracy is almost the same. "One against the others" was employed in the present study. The input class is compared to all other classes in this method (Araghinejad 2013).

Methodology

In order to model water quality classification by means of PNN, SVM and KNN, a set of 172 water samples and their quality classes were firstly chosen as the model's inputs and outputs in a way that water quality parameters were considered as the model's input while the water quality classes obtained from NSFWQI were chosen as the model's output. 75% of datasets were chosen as the calibration dataset while the rest were considered for the testing set. In order to prevent the model's over training at the calibration phase, cross validation was applied. Regarding the cross validation process, each time that the model is run with a specific parameter, a datum from the training set is put aside as the validation datum and the model is trained based on the remaining data; the above-mentioned datum will be returned to the training set and another datum from the training data is put aside as the validation datum and the network will be trained again. The process will be reiterated up until all training data are used in the validation phase. Afterwards, appropriate parameters for the three applied models will be determined in accordance with the model's average error in validation data estimation. Figure 2 illustrates different stages of the present research study.

Assessment criteria

The assessment criteria of error rate (ER), error value (EV) and accuracy (Acc) were applied at testing and calibration phases in order to assess the performance of the employed models. Error rate reveals the number of misclassified data while error value shows the magnitude of classification error in misclassified data. It is taken for granted that zero represents no errors in the classification made by these two criteria while the values higher than zero represent classification error. Accuracy illustrates the percentage of correct classification. These criteria are defined as follows:

$$ER = \frac{a}{a+b} \times 100, \tag{9}$$

$$EV = (observedclass - simulatedclass)^2, \tag{10}$$

$$Acc = \frac{b}{a+b} \times 100, \tag{11}$$

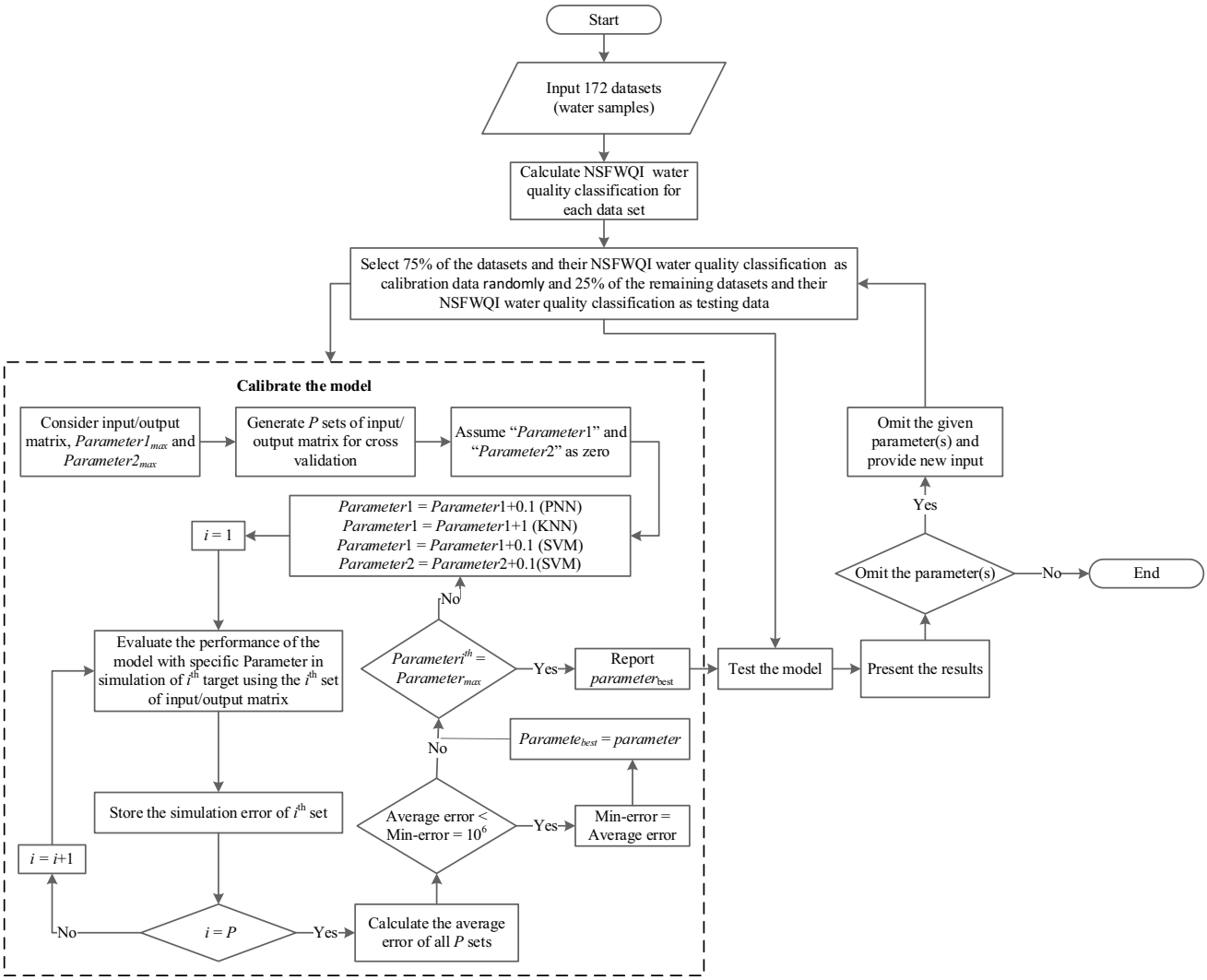

Fig. 2 Flowchart of the applied process where Parameter 1 is Spread for PNN, Parameter 1 is number of neighbors for KNN and Parameter 1 and Parameter 2 is C and γ for SVM

where a shows the classes predicted mistakenly and b represents classes predicted correctly.

Determining parameters affecting water quality classification

In order to determine the effective parameters in rivers' water quality classification by means of the above-mentioned models, one parameter out of nine considered parameters, was firstly removed and the effect of this removal on the process of water quality classification was studied. Aiming at further analysis of the model, a combination of two, three and more parameters was removed and the subsequent performance of the model was studied by means of the above-mentioned criteria. The results are discussed below.

Results and discussion

As mentioned earlier in the present paper the water quality parameters were considered as the data-driven models' input while their corresponding water quality classes were regarded as models' output. Therefore, quality classifications of the studied samples were determined using NSFWQI. The results revealed that samples could be categorized into three classes of good, bad and medium. In this case three models including PNN, KNN and SVM were developed and analyzed and the results are presented below.

Results of PNN modeling

Appropriate training of PNN necessitates determination of the optimal parameter value relevant to this method. The optimal value of this parameter plays a significant role in

the performance of PNN model and prevention of over training. The optimal value of *Spread Parameter* was, therefore, determined by trial and error by using cross validation. Having determined the optimal model, it was subsequently used to determine the water quality class. The results obtained from water quality classification by PNN model at calibration and testing phases revealed that out of 129 and 43 data which were respectively used at the calibration and testing phases, 7 and 4 samples were misclassified with one class difference; the error values were, therefore, 7 and 4 at the calibration and testing phases, respectively. It bears the meaning that the model is capable of accurately classifying water quality without a need to calculate NSFWQI. The results obtained from assessment criteria revealed that PNN enjoyed the accuracy of 94.57 and 90.70% at the calibration and testing phases respectively. It could be regarded appropriate performance in classifying water quality. Figure 3 sums up the results obtained from the analysis of the model at calibration and testing phases.

In order to determine the effective parameters in rivers' water classification by means of PNN model, one parameter out of nine considered parameters was firstly omitted and the effect of this removal on the process of predicting water classification was studied. Table 4 shows the results obtained subsequent to the removal of one parameter. As it could be seen, in case total solids and turbidity parameters are removed, the assessment criterion of error rate goes up from 5.43 to 7.75% at the calibration phases while the classification accuracy falls down from 94.57 to 92.25%. Omitting parameter fecal coliform also resulted in an enormous increase in classification error so that error rates at calibration and testing phases went up to 17.83 and 23.26% respectively while the classification accuracy fell down to 82.17 and 76.74% at the calibration and testing phases respectively. Furthermore, the removal of this parameter resulted in a condition in which 23 and 10 samples out of 129 and 43 assigned to calibration and testing were misclassified with one class difference. Aiming at further analysis of the model, a combination of two parameters and then three and so on were omitted. Table 5 indicates the performance of PNN subsequent to combinative removal of some parameters. Having compared the assessment criteria obtained from removal of different water quality parameters, it was revealed that fecal coliform, total solids and turbidity are the only parameters affecting water quality classification in case PNN model is employed. As it could be seen in Table 6, if pH, total phosphates, temperature, BOD, DO and nitrate are removed, the model experiences no changes in terms of water quality classification at calibration and testing phase with error rates of 5.43 and 9.30%, error values of 7 and 4 and accuracy of 94.57 and 90.70% respectively.

The results obtained from removal of water quality parameters in PNN model revealed that removal of total solids and turbidity parameters reduces the classification accuracy by 2.46%. Moreover, removal of fecal coliform parameter resulted in reduction of accuracy by 13.11 and 18.85% at calibration and testing phases respectively. Fecal coliform was, therefore, chosen as the most effective parameter in water quality classification in this model since if omitted, the error rate in water quality classification will be dramatically increased. Given the fact that annual cost of monitoring in order to determine water quality parameters

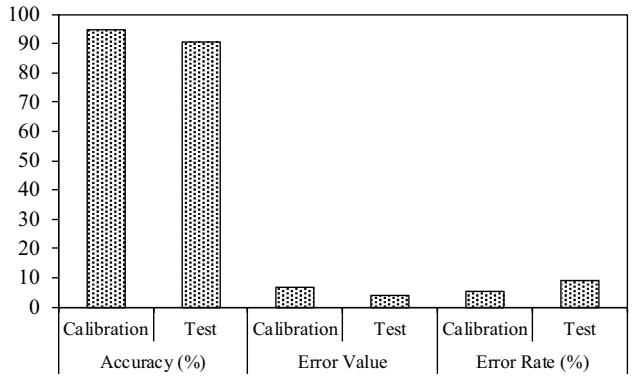

Fig. 3 The results obtained from the analysis of the PNN model at calibration and testing phases

Table 4 The effects of omitting a single parameter on the classification of water quality made by PNN model

Omitted parameter	Accuracy (%)		Error value		Error rate (%)	
	Calibration	Test	Calibration	Test	Calibration	Test
pH	94.57	90.70	7	4	5.43	9.30
Turb	92.25	90.70	10	4	7.75	9.30
TP	94.57	90.70	7	4	5.43	9.30
Temp	94.57	90.70	7	4	5.43	9.30
BOD	94.57	90.70	7	4	5.43	9.30
DO	94.57	90.70	7	4	5.43	9.30
Nitrate	94.57	90.70	7	4	5.43	9.30
FC	82.17	76.74	23	10	17.83	23.26
TS	92.25	90.70	10	4	7.75	9.30

Table 5 The effects of omitting more than one parameter on the classification of water quality made by PNN model

Omitted parameters	Accuracy (%)		Error value		Error Rate (%)	
	Calibration	Test	Calibration	Test	Calibration	Test
Turb, pH	92.25	90.70	10	4	7.75	9.30
Turb, FC	80.62	79.07	25	9	20.93	19.38
DO, TP	94.57	90.70	7	4	5.43	9.30
BOD, temp	94.57	90.70	7	4	5.43	9.30
FC, TS	82.17	76.74	23	10	17.83	23.26
DO, nitrate	94.57	90.70	7	4	5.43	9.30
Turb, TP, pH	92.25	90.70	10	4	7.75	9.30
Turb, BOD, TS	89.92	90.70	13	4	10.08	9.30
FC, temp, DO	82.17	76.74	23	10	17.83	23.26
Nitrate, DO, BOD, TP, temp, pH	94.57	90.70	7	4	5.43	9.30

Table 6 The effects of omitting a single parameter on the classification of water quality made by KNN model

Omitted parameter	Accuracy (%)		Error value		Error rate (%)	
	Calibration	Test	Calibration	Test	Calibration	Test
pH	89.15	90.70	14	4	10.85	9.30
Turb	89.15	90.70	14	4	10.85	9.30
TP	89.15	90.70	14	4	10.85	9.30
Temp	89.15	90.70	14	4	10.85	9.30
BOD	89.15	90.70	14	4	10.85	9.30
DO	89.15	90.70	14	4	10.85	9.30
Nitrate	89.15	90.70	14	4	10.85	9.30
FC	83.72	81.40	21	8	16.28	18.60
TS	92.25	90.70	10	4	7.75	9.30

is of paramount importance, PNN is capable of classifying water quality by three water quality parameters of fecal coliform, total solids and turbidity. It is, therefore, recommended that if water quality is going to be classified by PNN model, parameters of fecal coliform, total solids and turbidity be measured monthly while other parameters be measured every few months.

Results of KNN modeling

The parameter that needs to be determined for proper running of the model in KNN is the number of neighbors. The optimal value of this parameter is calculated for KNN model through applying cross-validation. The results obtained from water quality classification through applying KNN model at calibration and testing phases revealed that out of 129 and 43 samples that were considered for calibration and testing phases respectively, 14 and 4 samples were misclassified with one class difference. The results obtained from assessment criteria revealed that KNN enjoys the accuracy of 89.15 and 90.70% and error rate of 10.85 and 9.30% at calibration and testing phases respectively. Figure 4 sums up the results obtained from the analysis of the model at calibration and testing phases.

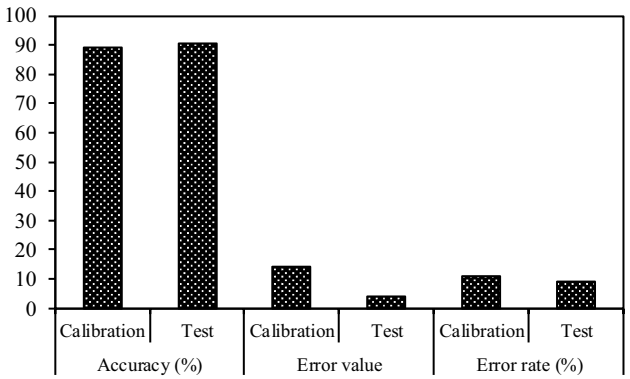

Fig. 4 The results obtained from assessment criteria at KNN calibration and testing phases

Table 6 shows the results obtained from omitting a single parameter in KNN model. As it could be seen, in case fecal coliform is removed, the error rate goes up from 10.853% and 9.302–16.28% and 18.60% while the classification accuracy falls down by 6.09 and 10.26% at calibration and

testing phases respectively. Furthermore, removal of total solids parameter led to an increase in classification accuracy from 89.15 to 92.25% at the calibration phase. In other words, the classification was improved. As it could be seen in Table 6, omitting other parameters did not bring about any changes in water quality classification. Aiming at further analysis of KNN model, combinations of two, three and more parameters were omitted. Table 7 shows the results obtained from combinative removal of some parameters for illustration. As it could be seen in Table 7, total solids is a parameter of paramount importance in classification so that if removed along with nitrate parameter, it results in classification error (accuracy may decrease by 2.63% at the testing phase) while if removed along with other parameters, it leads to an improvement in assessment criteria at the calibration phase. Subsequent to trial and error and removal of all combinations of two, three and multiple parameters, it was revealed that the model represents its best performance in estimation of water quality classification at the accuracy of 92.25% and 90.70% and error rates of 7.75% and 9.30% at calibration and testing phases respectively under the condition that three parameters of turbidity, nitrate and fecal coliform are used. Compared to the condition in which no parameter is removed, classification accuracy increased by 3.48% under the condition stipulated above. Fecal coliform was the parameter of utmost importance in quality classification in this model as well so that its removal decreased classification accuracy by 10.26%.

Results of SVM modeling

Regarding SVM, the parameters C (in Eq. 5) and γ (in Eq. 7) need to be determined by the user through trial and error. Therefore, the best values of these two parameters were calculated through cross validation and trial and error under two conditions: (1) omitting water quality parameter, (2) without omitting water quality parameter. Figure 5 shows the results obtained from water quality classification of Karoon

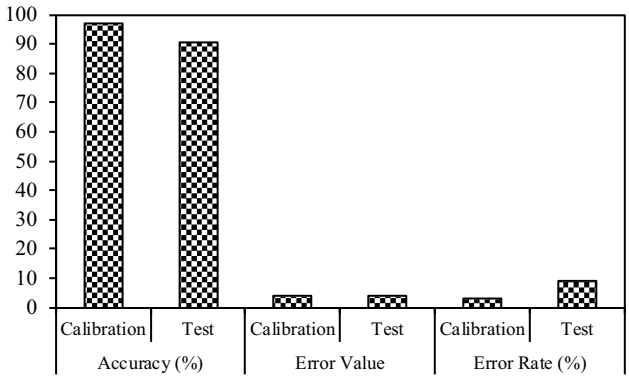

Fig. 5 The results obtained from assessment criteria at SVM calibration and testing phases

River through applying SVM model. The obtained accuracy was 96.90 and 90.70% at calibration and testing phases respectively. Moreover, it was revealed that out of 129 and 43 samples considered for calibration and testing phases, 4 and 4 samples were misclassified with one class difference.

Table 8 shows the results obtained from omitting a single parameter when SVM model is applied. As it could be seen in Table 8, the removal of a single parameter has led to a decrease in accuracy and an increase in classification error rate. The highest increase in error rate was observed at the testing phase subsequent to omitting total phosphates, DO and BOD parameters and the error rate stood at 125.01%, 150.01% and 100.01% respectively. On the other hand, omitting nitrate and turbidity parameters resulted in the lowest error rate. Aiming at further analysis, all possible combinations of two and three parameters were omitted and the subsequent performance of SVM model in classification was studied. Table 9 shows some results obtained from performance of SVM model subsequent to omitting some combinations of two, three or more parameters. Regarding this model, it was revealed as well that subsequent to trial and error and omitting all possible combinations of two, three

Table 7 The effects of omitting more than one parameter on the classification of water quality made by KNN model

Omitted parameters	Accuracy (%)		Error value		Error rate (%)	
	Calibration	Test	Calibration	Test	Calibration	Test
TS, pH	92.25	90.70	10	4	7.75	9.30
TP, DO	89.15	90.70	14	4	10.85	9.30
Temp, BOD	89.15	90.70	14	4	10.85	9.30
Nitrate, TS	89.15	88.37	14	5	10.85	11.63
Turb, temp, TS	90.70	90.70	12	4	9.30	9.30
pH, TP, TS	92.25	90.70	10	4	7.75	9.30
TP, nitrate, TS	89.15	88.37	14	5	10.85	11.63
pH, turb, TP, temp, BOD, DO	89.15	90.70	14	4	10.85	9.30
pH, turb, TP, temp, BOD, DO, TS	90.70	90.70	12	4	9.30	9.30
pH, TP, temp, BOD, DO, TS	92.25	90.70	10	4	7.75	9.30

Table 8 The effects of omitting a single parameter on the classification of water quality made by SVM model

Omitted parameter	Accuracy (%)		Error value		Error rate (%)	
	Calibration	Test	Calibration	Test	Calibration	Test
pH	97.67	83.72	7	3	2.33	16.28
Turb	96.12	90.70	5	7	3.88	9.30
TP	96.12	79.07	5	15	3.88	20.93
Temp	98.45	88.37	4	8	1.16	11.63
BOD	97.67	81.40	3	8	2.33	18.61
DO	99.23	76.74	1	16	0.78	23.26
Nitrate	96.12	90.70	5	4	3.88	9.30
FC	95.35	88.37	6	8	4.65	11.63
TS	92.25	88.37	10	5	7.75	11.63

Table 9 The effects of omitting more than one parameter on the classification of water quality made by SVM model

Omitted parameters	Accuracy (%)		Error value		Error rate (%)	
	Calibration	Test	Calibration	Test	Calibration	Test
TP, DO	88.37	69.77	15	19	11.63	30.23
TS, pH	93.02	90.70	9	4	6.98	9.30
Turb, nitrate	94.57	93.02	7	3	5.43	6.98
TS, pH, turb	82.17	83.72	23	7	17.83	16.28
Turb, TP, DO	95.35	65.12	6	21	4.65	34.88
TP, BOD, FC	92.25	60.47	10	17	7.75	39.54
Temp, BOD, TS	96.90	76.74	4	10	4.65	23.26
pH, turb, TP, temp, TS	82.95	83.72	22	7	17.05	16.28
pH, turb, TP, temp, TS, nitrate	78.30	83.72	31	7	21.71	16.28
pH, turb, temp, nitrate, TS	88.37	88.37	15	5	11.63	11.63

or more parameters, it is possible to observe the best performance of SVM model with the fewest water quality parameters through measuring total phosphates, BOD, DO and fecal coliform parameters. As it could be seen in Table 9, the water quality of Karoon River could be classified by the above-mentioned parameters at 88.37% accuracy and 11.63% error rate at calibration and testing phases. Furthermore, the error value assessment criteria revealed that 5 and 15 parameters were misclassified with one class difference.

Discussion

Figure 6 shows the results obtained from comparing the three models. As it could be seen, under the condition that no parameter is omitted, all the three models showed the same performance with the accuracy of 90.7%, error rate of 9.30% and error value of 4 at the testing phase while SVM model with the accuracy of 96.90%, error rate of 3.10% and error value of 4 delivered the best performance at the calibration phase.

Figure 7 shows the best water quality classification results obtained subsequent to omitting water quality parameters. Under this circumstance, having measured merely three parameters of total solids, fecal coliform and turbidity at 94.57 and 90.70% accuracy and 5.43 and 9.30% error rate at the calibration and testing phases respectively, PNN model delivered the best performance. The worst results were obtained from SVM model that (by measuring four parameters of total phosphates, DO, BOD and fecal coliform) had the highest error with 88.37% accuracy and 11.63% error rate at calibration and testing phases.

Having compared the effects of omitting single parameters, it was concluded that omitting fecal coliform parameter in PNN and KNN models results in major classification error. On the other hand, regarding SVM model, omitting two parameters of total phosphates and DO resulted in relatively high classification error. It is, therefore, recommended that fecal coliform in KNN and PNN models and total phosphates and DO parameters in SVM model be constantly measured for determining water quality classification in Karoon river. It is worth noting that compared to other two models; SVM model's training process and the optimum value of C parameter are more difficult and time-consuming. Given the fact the training process of the other two models is faster than that of SVM model; it is preferred to apply PNN and KNN models in Karoon River's water quality classification.

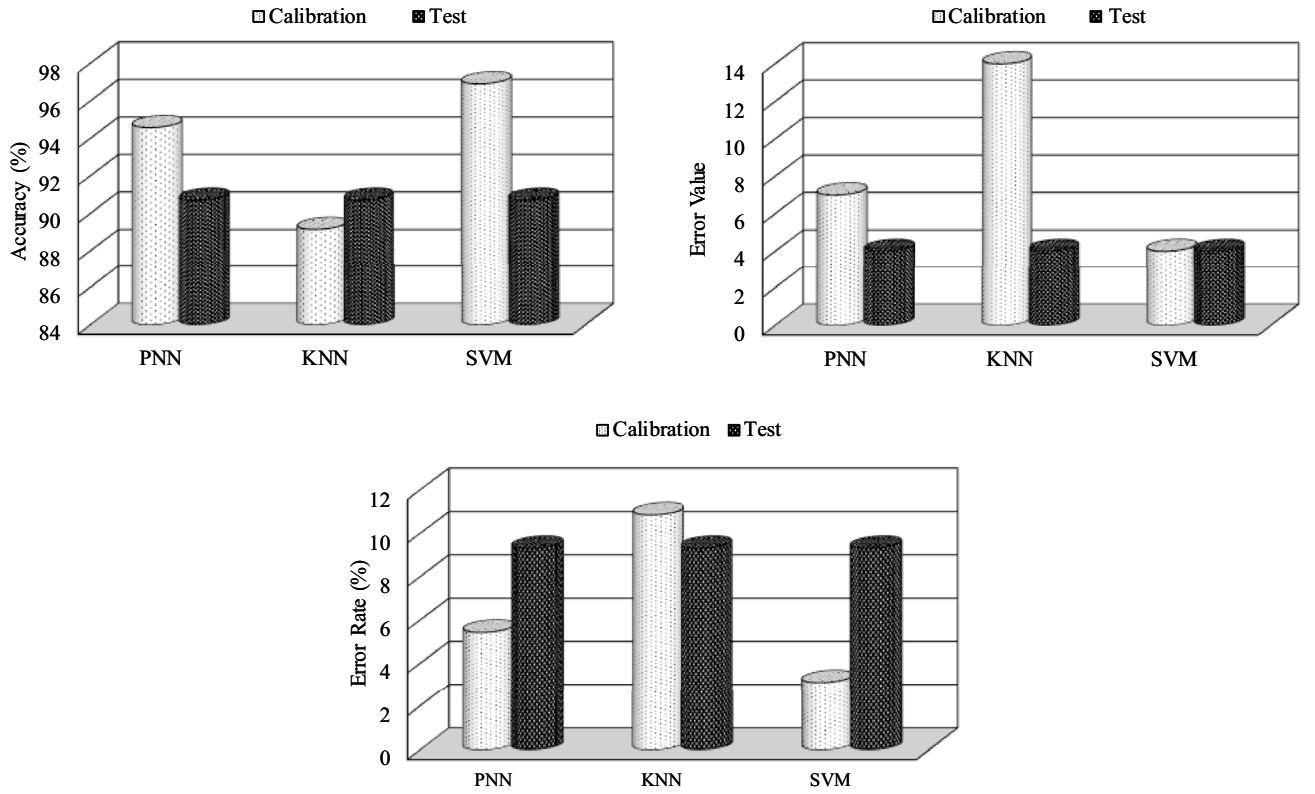

Fig. 6 Comparing PNN, SVM and KNN models in terms of water quality classification on Karoon (without omitting parameters)

Generally, the results indicated that PNN could be regarded a proper alternative to water quality indices in order to classify water quality and assess water quality parameter. As it was said, PNN is a supervised model. One of the most prominent features of these models is that the required network size and the classification error rate are directly embodied in the network's structure process; therefore, the proposed model often has a small network structure and delivers acceptable classification accuracy (Mao et al. 2000). Xue et al. (2005) concluded in their study on classifying the active compounds in medical plants that compared to back-propagation (BP) neural networks, PNN enjoys a faster and simpler structure since it has only one adjustable parameter; it could be, therefore, considered a proper prediction and classification method. Belayneh et al. (2014) claimed in their study on drought prediction that although SVM theoretically delivers a better performance than ANN since instead of applying empirical risk minimization principle, it applies structural risk minimization principle; it might deliver a performance the same as or even worse than that of ANN. The results obtained in the present study are in line with their results. Modaresi and Araghinejad (2014) concluded in their study that PNN could be regarded a proper water quality classification method enjoying the advantage of fast training process. In order to simulate TSD and EC parameters by other water quality parameters' values, Khaki et al. (2015) applied artificial neural networks (ANN) and neuro-fuzzy system and concluded that artificial intelligence approaches and neuro-fuzzy systems are capable of interpreting water quality parameters' behavior.

Despite the fact that the mentioned approach could determine water quality classification data over the periods when sampling has been done merely by some parameters, it should be noted that performing this research in developing countries is always faced with limitations such as lack of in situ data. The limitation of water quality data is more severe than quantitative one; moreover, the length of time series of water quality samples and the number of measured parameters could be regarded among these limitations. Although the maximum available water quality data was used in the present study, the number of water quality samples is small for accurate modeling. Therefore, the mentioned approach is suggested to be evaluated in areas having more in situ data.

Conclusions

In the present study, 172 water samples taken from Karoon River at eight stations through applying NSFWQI methods and SVM, KNN and PNN models were classified and

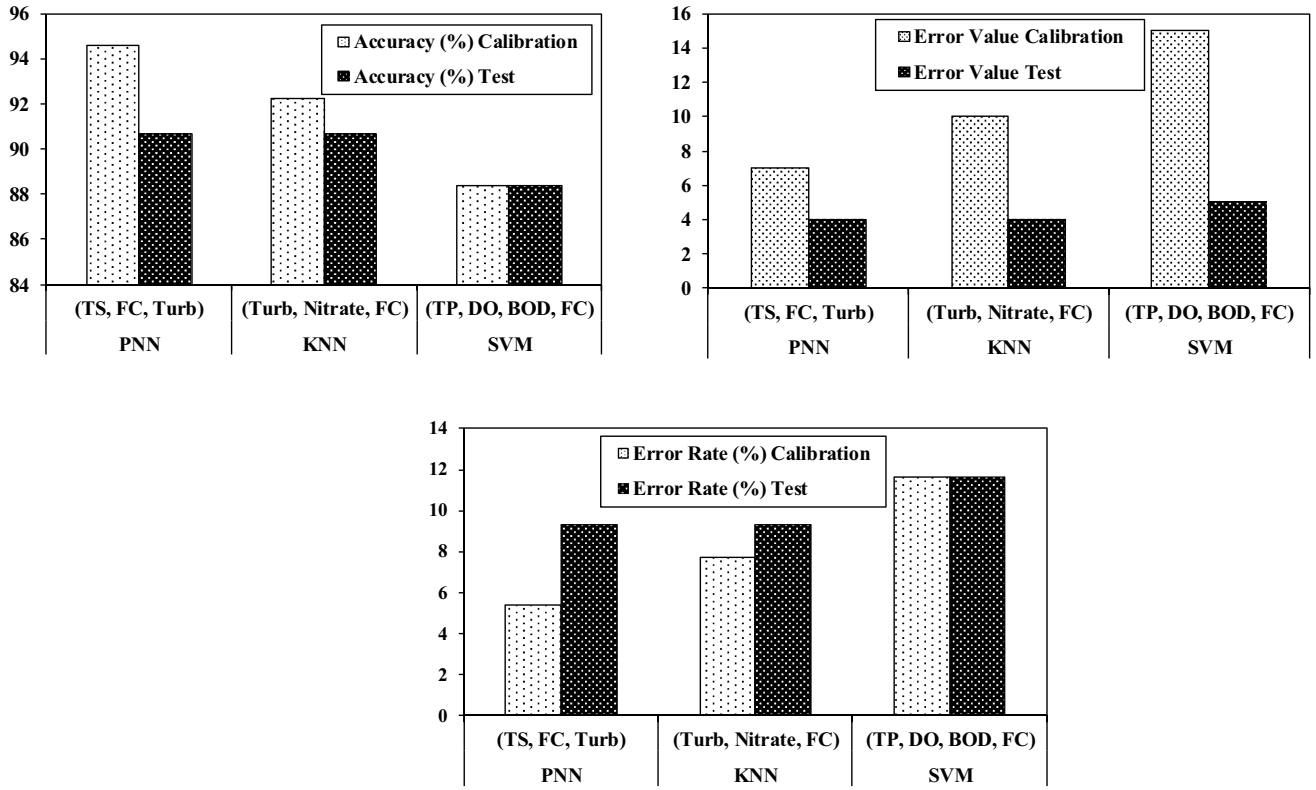

Fig. 7 Comparing PNN, SVM and KNN models in terms of water quality classification on Karoon (with omitting parameters)

compared in order to analyze the surface water quality assessment methods. Given the fact that water quality classification by means of water quality indices is time-consuming and costly, data-driven methods were proposed in the present study as an alternative to classify water quality. The results obtained from NSFWQI revealed that samples could be classified into three categories of good, bad and medium. Making use of three water quality parameters of turbidity, fecal coliform and total solids, PNN model with the accuracy of 94.57 and 90.70% at the calibration and testing phases respectively delivered the most reasonable performance compared to other two models. PNN model may, therefore, reduce the sampling costs and computation time for water quality classification and it could be regarded an appropriate alternative to NSFWQI. Moreover, the results obtained from the comparisons made among the three models revealed that fecal coliform is the most effective parameter in water quality classification since if omitted, a great error might occur in water quality classification. It is recommended that other models be employed for water quality classification in future studies in order to determine the water quality parameters which might have an effect on water quality classification.

Acknowledgements The authors are grateful to the University of Tehran for providing all required facilities to do the present study and its papers.

References

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

Aggarwal SK, Goel A, Singh VP (2012) Stage and discharge forecasting by SVM and ANN techniques. Water Resour Manag 26(13):3705–3724. https://doi.org/10.1007/s11269-012-0098-x

Araghinejad S (2013) Data-driven modeling: using MATLAB® in water resources and environmental engineering, vol 67. Springer Science & Business Media, Berlin

Barzegar R, Sattarpour M, Nikudel MR, Moghaddam AA (2016) Comparative evaluation of artificial intelligence models for prediction of uniaxial compressive strength of travertine rocks, case study: Azarshahr area, NW Iran. Model Earth Syst Environ 2(2):76. https://doi.org/10.1007/s40808-016-0132-8

Belayneh A, Adamowski J, Khalil B, Ozga-Zielinski B (2014) Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. J Hydrol 508:418–429. https://doi.org/10.1016/j.jhydrol.2013.10.052

Brown RM, McClelland NI, Deininger RA, Tozer RG (1970) A water quality index—do we dare. Water Sew Works 117:339–343

Brown RM, McClelland NI, Deininger RA, O'Connor MF (1972) A water quality index—crashing the psychological barrier. In: Indicators of environmental quality. Springer, New York, pp 173–182

Cacoullos T (1966) Estimation of a multivariate density. Ann Inst Stat Math 18(1):179–189

Chen AS, Leung MT, Daouk H (2003) Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. Comput Oper Res 30(6):901–923

Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw 17(1):113–126. https://doi.org/10.1016/S0893-6080(03)00169-2

Chomboon K, Chujai P, Teerarassamee P, Kerdprasop K, Kerdprasop N (2015) An empirical study of distance metrics for k-nearest neighbor algorithm. In: Proceedings of the 3rd international conference on industrial application engineering, Kitakyushu, Japan, March 28–31, 2015. ICIAE2015

Ge SS, Yang Y, Lee TH (2008) Hand gesture recognition and tracking based on distributed locally linear embedding. Image Vis Comput 26(12):1607–1620. https://doi.org/10.1016/j.imavis.2008.03.004

Genç O, Dağ A (2016) A machine learning-based approach to predict the velocity profiles in small streams. Water Resour Manage 30(1):43–61. https://doi.org/10.1007/s11269-015-1123-7

Hajmeer M, Basheer I (2002) A probabilistic neural network approach for modeling and classification of bacterial growth/no-growth data. J Microbiol Methods 51(2):217–226. https://doi.org/10.1016/S0167-7012(02)00080-5

Hipni A, El-shafie A, Najah A, Karim OA, Hussain A, Mukhlisin M (2013) Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). Water Resour Manag 27(10):3803–3823. https://doi.org/10.1007/s11269-013-0382-4

Hodge V, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22(2):85–126. https://doi.org/10.1007/s10462-004-4304-y

Horton RK (1965) An index number system for rating water quality. J Water Pollut Control Fed 37(3):300–306

Hosseini-Moghari SM, Ebrahimi K, Azarnivand A (2015) Groundwater quality assessment with respect to fuzzy water quality index (FWQI): an application of expert systems in environmental monitoring. Environ Earth Sci 74(10):7229–7238. https://doi.org/10.1007/s12665-015-4703-1

Hosseini-Moghari SM, Araghinejad S, Azarnivand A (2017) Drought forecasting using data-driven methods and an evolutionary algorithm. Model Earth Syst Environ 1–15

Javan K, Lialestani MRFH., Nejadhossein M (2015) A comparison of ANN and HSPF models for runoff simulation in Gharehsoo River watershed, Iran. Model Earth Syst Environ 1(4):41. https://doi.org/10.1007/s40808-015-0042-1

Jung NC, Popescu I, Kelderman P, Solomatine DP, Price RK (2010) Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea. J Hydroinformatics 12(3):262–274. https://doi.org/10.2166/hydro.2009.004

Kasiviswanathan KS, Sudheer KP (2016) Comparison of methods used for quantifying prediction interval in artificial neural network hydrologic models. Model Earth Syst Environ 2(1):22. https://doi.org/10.1007/s40808-016-0079-9

Khaki M, Yusoff I, Islami N (2015) Application of the artificial neural network and neuro-fuzzy system for assessment of groundwater quality. CLEAN Soil Air Water 43(4):551–560. https://doi.org/10.1002/clen.201400267

Kim DK, Lee JJ, Lee JH, Chang SK (2005) Application of probabilistic neural networks for prediction of concrete strength. J Mater Civ Eng 17(3):353–362. https://doi.org/10.1061/(ASCE)0899-1561(2005)17:3(353)

Lan Y (2014) Forecasting performance of support vector machine for the Poyang Lake's water level. Water Sci Technol 70(9):1488–1495. https://doi.org/10.2166/wst.2014.396

Landwehr JM, Deininger RA (1976) A comparison of several water quality indexes. J (Water Pollut Control Fed) 954–958

Li Z, Zhou M, Xu L, Lin H, Pu H (2014) Training sparse SVM on the core sets of fitting-planes. Neurocomputing 130:20–27. https://doi.org/10.1016/j.neucom.2013.04.046

Li X, Sha J, Wang ZL (2017) A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. Hydrol Res 48(5):1214–1225

Lin JY, Cheng CT, Chau KW (2006) Using support vector machines for long-term discharge prediction. Hydrol Sci J 51(4):599–612. https://doi.org/10.1623/hysj.51.4.599

Liu M, Lu J (2014) Support vector machine——an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? Environ Sci Pollut Res 21(18):11036–11053. https://doi.org/10.1007/s11356-014-3046-x

Liu Y, Wang H, Zhang H, Liber K (2016) A comprehensive support vector machine-based classification model for soil quality assessment. Soil Tillage Res 155:19–26. https://doi.org/10.1016/j.still.2015.07.006

Londhe S, Panse-Aglave G (2015) Modelling stage–discharge relationship using data-driven techniques. ISH J Hydraul Eng 21(2):207–215. https://doi.org/10.1080/09715010.2015.1007092

Mahmoudi N, Orouji H, Fallah-Mehdipour E (2016) Integration of shuffled frog leaping algorithm and support vector regression for prediction of water quality parameters. Water Resour Manage 30(7):2195–2211. https://doi.org/10.1007/s11269-016-1280-3

Mao KZ, Tan KC, Ser W (2000) Probabilistic neural-network structure determination for pattern classification. IEEE Trans Neural Netw 11(4):1009–1016. https://doi.org/10.1109/72.857781

Martins S, Bernardo N, Ogashawara I, Alcantara E (2016) Support vector machine algorithm optimal parameterization for change detection mapping in funil hydroelectric reservoir (Rio de Janeiro State, Brazil). Model Earth Syst Environ 2(3):138. https://doi.org/10.1007/s40808-016-0190-y

Masters T (1995) Advanced algorithms for neural networks: a C + + sourcebook. Wiley, New York

Modaresi F, Araghinejad S (2014) A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water Resour Manage 28(12):4095–4111. https://doi.org/10.1007/s11269-014-0730-z

Mohammadi Ghaleni M, Ebrahimi K (2015) Effects of human activities and climate variability on water resources in the Saveh plain. Iran Environ Monit Assess 187(2):35. https://doi.org/10.1007/s10661-014-4243-2

Mohammadpour R, Shaharuddin S, Zakaria NA, Ghani AA, Vakili M, Chan NW (2016) Prediction of water quality index in free surface constructed wetlands. Environ Earth Sci 75(2):139. https://doi.org/10.1007/s12665-015-4905-6

Parsaie A, Yonesi HA, Najafian S (2015) Predictive modeling of discharge in compound open channel by support vector machine technique. Model Earth Syst Environ 1(1–2):1. https://doi.org/10.1007/s40808-015-0002-9

Rahimi J, Khalili A, Bazrafshan J (2017) Analysis of late spring frost dates over Iran under current climate and future scenarios. Model Earth Syst Environ 1–10

Sadat-Noori SM, Ebrahimi K, Liaghat AM (2014) Groundwater quality assessment using the water quality index and GIS in Saveh–Nobaran aquifer. Iran Environ Earth Sci 71(9):3827–3843. https://doi.org/10.1007/s12665-013-2770-8

Sakizadeh M (2015) Assessment the performance of classification methods in water quality studies, a case study in Karaj River.

Environ Monit Assess 187(9):573. https://doi.org/10.1007/s10661-015-4761-6
Sakizadeh M (2016) Artificial intelligence for the prediction of water quality index in groundwater systems. Model Earth Syst Environ 2(1):8
Sakizadeh M, Mirzaei R (2016) A comparative study of performance of K-nearest neighbors and support vector machines for classification of groundwater. J Mining Environ 7(2):149–164. https://doi.org/10.22044/jme.2016.480
Sharip Z, Saman JM, Noordin N, Majizat A, Suratman S, Shaaban AJ (2016) Assessing the spatial water quality dynamics in Putrajaya Lake: a modelling approach. Model Earth Syst Environ 2(1):46
Sharma N, Zakaullah M, Tiwari H, Kumar D (2015) Runoff and sediment yield modeling using ANN and support vector machines: a case study from Nepal watershed. Model Earth Syst Environ 1(3):23. https://doi.org/10.1007/s40808-015-0027-0
Shirmohammadi B, Vafakhah M, Moosavi V, Moghaddamnia A (2013) Application of several data-driven techniques for predicting groundwater level. Water Resour Manage 27(2):419–432. https://doi.org/10.1007/s11269-012-0194-y
Specht DF (1990) Probabilistic neural networks. Neural Netw 3(1):109–118. https://doi.org/10.1016/0893-6080(90)90049-Q. doi
Sudheer KP, Gosain AK, Ramasastri KS (2002) A data-driven algorithm for constructing artificial neural network rainfall-runoff models. Hydrol Process 16(6):1325–1330. https://doi.org/10.1002/hyp.554
Tian J, Li C, Liu J, Yu F, Cheng S, Zhao N, Wan Jaafar WZ (2016) Groundwater depth prediction using data-driven models with the assistance of Gamma test. Sustainability 8(11):1076. https://doi.org/10.3390/su8111076
Tokar AS, Markus M (2000) Precipitation-runoff modeling using artificial neural networks and conceptual models. J Hydrol Eng 5(2):156–161. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(156)
Towler E, Rajagopalan B, Seidel C, Summers RS (2009) Simulating ensembles of source water quality using a K-nearest neighbor resampling approach. Environ Sci Technol 43(5):1407–1411. https://doi.org/10.1021/es8021182
Vapnik V (1995) The nature of statistical learning theory. Springer, New York
Walker D, Jakovljević D, Savić D, Radovanović M (2015) Multicriterion water quality analysis of the Danube River in Serbia: a visualisation approach. Water Res 79:158–172. https://doi.org/10.1016/j.watres.2015.03.020
Wasserman PD (1993) Advanced methods in neural computing. Wiley, New York
Weston J, Watkins C (1998) Multi-class support vector machines. Technical Report, University of London
Wu CL, Chau KW, Li YS (2009) Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resour Res 45(8). https://doi.org/10.1029/2007WR006737
Xue CX, Zhang XY, Liu MC, Hu ZD, Fan BT (2005) Study of probabilistic neural networks to classify the active compounds in medicinal plants. J Pharm Biomed Anal 38(3):497–507. https://doi.org/10.1016/j.jpba.2005.01.035
Yoon H, Hyun Y, Ha K, Lee KK, Kim GB (2016) A method to improve the stability and accuracy of ANN-and SVM-based time series models for long-term groundwater level predictions. Comput Geosci 90:144–155. https://doi.org/10.1016/j.cageo.2016.03.002