JME

Journal of Mining & Environment,
Vol.7, No.2, 2016, 149-164.

A comparative study of performance of K-nearest neighbors and support vector machines for classification of groundwater

M. Sakizadeh[1]* and R. Mirzaei[2]

1. Department of Environmental Sciences, Faculty of Sciences, Shahid Rajaee Teacher Training University, Tehran, Iran
2. Department of Environmental Sciences,University of Kashan, Kashan, Iran

Received 17 June 2015; received in revised form 23 July 2015; accepted 30 August 2015
**Corresponding author: msakizadeh@gmail.com (M. Sakizadeh).*

Abstract
The aim of this work is to examine the feasibilities of the support vector machines (SVMs) and K-nearest neighbor (K-NN) classifier methods for the classification of an aquifer in the Khuzestan Province, Iran. For this purpose, 17 groundwater quality variables including EC, TDS, turbidity, pH, total hardness, Ca, Mg, total alkalinity, sulfate, nitrate, nitrite, fluoride, phosphate, Fe, Mn, Cu, and Cr(VI) from 41 wells and springs were used during an eight-year time period (2006 to 2013). The cluster analysis was used, leading to a dendrogram that differentiated two distinct groups. The factor analysis extracted eight factors accumulatively, accounting for 90.97% of the total variance. Thus the variations in 17 variables could be covered by just eight factors. K-NN and SVMs were applied for the classification of the aquifer under study. The results of SVMs indicated that the best performed model was related to an exponent of degree one with an accuracy of 94% for the test data set, in which the sensitivity and specificity were 1.00 and 0.87, respectively. In addition, there was no significant difference among the results of different kernels, indicating that an acceptable result can be achieved by selecting the optimum parameters for a kernel. The results of K-NN showed roughly a lower efficiency compared with those of SVMs, where the sensitivity and specificity was reduced to 0.90 and 0.88, respectively, although the accuracy of the model was 93%. A sensitivity analysis was performed on the groundwater quality variables, suggesting that calcium next to nitrate were the most influential parameters in the classification of this aquifer.

Keywords: *Groundwater, Support Vector Machines, K-Nearest Neighbors, Kernel Functions.*

1. Introduction

During the last few decades, a rapid population growth has put pressure on the groundwater resources in Iran, especially in the arid and semi-arid areas, where the surface water resources cannot obviate the requirements of the people for domestic, industrial, and agricultural activities. The over-exploitation of groundwater resources besides the recent drought in the regions that began from 2007 resulted in a precipitous depletion of this valuable resource [1, 2]. In addition, the quality of the water resources has aggravated in the recent years [3, 4]. In this regard, the groundwater quality classification is a tool for the local managers to use in land-use management decisions. For instance, one of the possible applications of aquifer classification is to locate activities in the areas where groundwater is already poor [5]. A classification task usually involves separating data into the training and testing sets. Each instance in the training set contains one "target value" (i.e. the quality rating in this case) and several "attributes" (i.e. groundwater quality variables). There are many classification methods that can be applied for this purpose such as K-nearest neighbor (K-NN) [6], support vector machines (SVMs) [7], discriminant analysis [8], classification trees [9], and probabilistic neural networks [10].

Among these methods, SVM is one the most recently applied classification methods in environmental researches [e.g. 11-13]. SVMs, which is based upon the structural risk

minimization (SRM) principle [14], seems to be a promising method for data mining, and have been used for both the classification and regression problems. The goal of SVMs is to produce a model (based on the training data) that predicts the target values of the test data, given only the test data attributes [15]. SVMs, essentially a kernel-based procedure, creates very competitive results with the best accessible classification methods, and needs only the smallest amounts of model tuning [16]. To the contrary, as mentioned by Akay [17], the type of kernel function, the optimum number of input features for SVM, and how to tune kernel parameters to reach the best generalization are the three problems during the SVMs model development.

On the other hand, as explained by Rokach [18], the nearest neighbor classifier method has many advantages over other ones. For instance, the generalization ability of a relatively small amount of data set is better in comparison with other classifiers such as the decision trees or neural networks. Moreover, new information can be incrementally incorporated at runtime, a property it shares with neural networks. Consequently, the nearest neighbor classifier can achieve a performance that is competitive to more modern and complex methods such as decision trees and neural networks.

SVMs has been utilized earlier for the prediction of nitrate levels in groundwater [7] and groundwater level predictions [12]. However, to the best of our knowledge, there is just one published literature [10] on the usage of SVMs for the groundwater quality classification. Thus the objective of this work to examine the feasibility of the SVMs and K-NN classifiers for the classification of an aquifer in the Khuzestan province in Iran based on the pollution level.

2. Materials and method

2.1. Studied area

Andimeshk is located in the northern part of the Khuzestan Province, south of Iran, with an area of 3100 km^2 (Figure 1). According to the latest census by the Iranian Statistical Center in 2012, the total population of this city is 167126, among which, 128774 inhabitants live in the urban areas and 34985 live in the rural areas. The average annual precipitation in the region is about 353 mm, and the average exploitation from the groundwater resources was 133981 thousand cubic meters/year in 2012, according to the report prepared by Khuzestan Water and Power Authority. The major irrigated broad-acre crops grown in the region are wheat, barley, and maize, in addition to fruits, melons, watermelons, and vegetables such as tomatoes and cucumbers [19]. The average water-level fluctuations between the dry and wet seasons are very low (about 0.5-1 m) because of the continuous recharge with the Dez and Karkhe Rivers [20]. The groundwater resources in the region are used for both the drinking and irrigation purposes.

2.2. Data pretreatment and Cluster analysis

Seventeen groundwater quality variables including EC, TDS, Turbidity, pH, total hardness(TH), Ca, Mg, total alkalinity (TA), sulfate, nitrate, nitrite, fluoride, phosphate, Fe, Mn, Cu and Cr (VI), associated with 41 wells and springs in Andimeshk were utilized in this study during an eight-year time period (2006 to 2013). The descriptive statistics related to these parameters are presented in Table 1.

Since the original variables were in different units, the operations involving the trace of the covariance matrix had no meaning. Therefore, the solution was to make the variances the same (i.e. use standard units), making the covariance matrix into a correlation matrix. Standardization of the data (i.e. data having zero mean and unit standard deviation) was implemented according to the following equation:

$$Z_i = (x_i - \bar{x}_i) / s_i \quad (1)$$

where $\bar{x}_i$ and s_i are the mean and standard deviations of the observed variables, respectively. Cluster analysis was used to consider the similarity among the sampling stations. Dendrograms could be useful for classification of the aquifer by the respective classifiers. In the cluster analysis, the objects (e.g. sampling stations) are grouped based on the similarities within a class and dissimilarities among the different classes [21]. A distance measure is used to examine the similarity and/or dissimilarity among the objects of interest. The most prevalent distance measures are Euclidean and Manhattan [22]. In order to cluster the sampling stations with respect to the seventeen stated groundwater quality variables, the average linkage method with Manhattan distance measure was utilized to produce the resultant dendrogram using the MINITAB (R2013b) software.

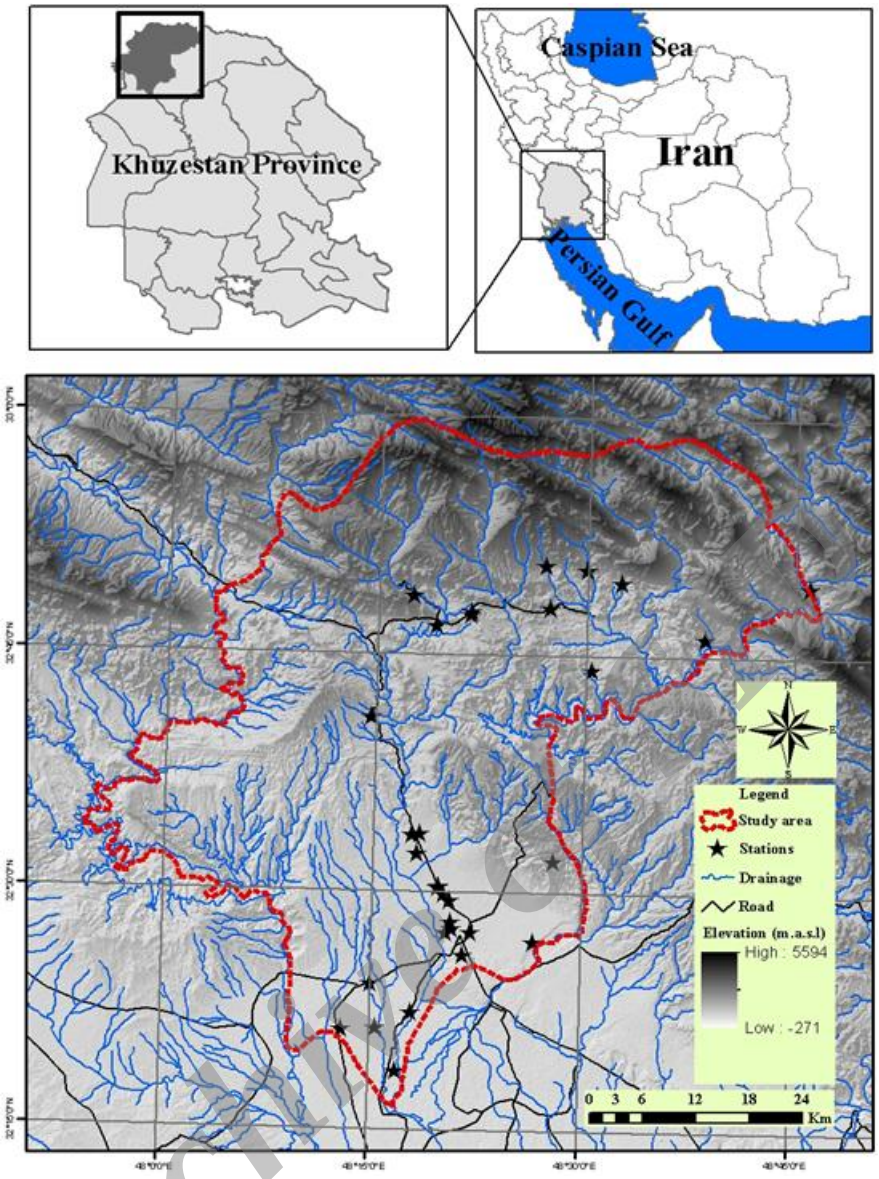

Figure 1. Location of studied area and sampling stations in Andimeshk aquifer,Iran.

Table 1. Descriptive statistics of groundwater quality data set used for classification.

Groundwater quality parameter	Average	Max.	Min.	Standard deviation
EC (µs/Cm)	552.68	850.00	270.00	161.23
TDS (mg/L)	291.35	439.23	130.00	82.70
Turbidity (NTU)	1.47	21.15	0.00	4.58
pH	7.68	8.14	6.22	0.34
Total hardness (mg/L)	246.09	349.46	95.76	63.67
Ca (mg/L)	64.24	105.65	22.40	17.76
Mg (mg/L)	20.39	35.17	7.77	7.10
Total alkalinity (mg/L)	124.36	240.80	48.00	41.39
Sulfate (mg/L)	95.68	193.75	7.00	54.82
Nitrate (mg/L)	26.68	84.90	5.05	18.51
Nitrite (mg/L)	0.02	0.06	0.00	0.01
Fluoride (mg/L)	0.37	0.86	0.17	0.17
Phosphate (mg/L)	0.14	0.24	0.03	0.05
Fe (mg/L)	0.03	0.09	0.01	0.02
Mn (mg/L)	0.05	0.15	0.00	0.03
Cu (mg/L	0.08	0.41	0.015	0.07
Cr (mg/L)	0.05	0.08	0.00	0.02

2.3. Factor analysis

In order to examine the suitability of the data for factor analysis/principal component analysis, the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests were performed. Factor analysis based on the principal component analysis was utilized to study the main factors responsible for the variation in the groundwater quality parameters in the studied area. The Vartimax rotation method was used to facilitate interpretation of the results obtained from the factor analysis.

2.4. Support Vector Machines (SVMs)

SVM is a supervised learning technique. In the linearly separable data set, all the patterns can be separated by a straight line or a hyper plane. In SVMs, it is implemented through maximization of the margin around a hyper plane that separates two classes by mapping the input space into a high dimensional space or feature space. The mapping is determined by a kernel function [13].

There are two stages in order to create and apply a SVMs classifier: training and prediction. In the training (or learning) stage, pairs of features and desired outputs (x_i, y_i) are given in order to design support vectors, which are used to predict the desired outputs. These support vectors constitute the prediction model. Later, after learning in the prediction phase, the prediction model is applied to predict outputs, y_i, for the previously unseen input feature vectors, $x_i = (x_1, x_2, ..., x_n)$. Let (w, b) denote the weight vector, $w = (w_1, w_2, ..., w_n)$, and bias, b, of the hyper plane that splits the data from both classes. At the training phase, the objective is to determine the separating hyper plane, later used to classify the unseen data [23].

The set of vectors is said to be optimally separated by the hyper plane if it is separated without error, and the distance between the closest vectors to the hyper plane is maximal. A separating hyper plane in the canonical form must satisfy the following constraints:

$$y^i\left[\langle w, x^i\rangle + b\right] \gg 1, i = 1, \ldots . l. \quad (2)$$

That is to say, in SVMs the training data points satisfying the constraints that $f(\mathbf{x}_i) = 1$ if $y_i = 1$, and $f(\mathbf{x}_i) = -1$ if $yi = -1$ are called the support vectors (SVs). In other words, the training points with non-zero weight are called the support vectors [24].

As a whole, as mentioned by Hsu et al. [15], given a training set of instance-label pairs $(\boldsymbol{x}_i, y_i), i=1,\ldots,l$ where $\boldsymbol{x}_i \in R^n$ and $y \in \{1, -1\}^l$, the support vector machine (SVM) require the solution to the following optimization problem:

$$\min_{w,b,\xi} {}^1\!/_2\, \boldsymbol{w}^T \boldsymbol{w} + C \sum_{i=1}^{l} \xi_i \quad (3)$$

$$\text{subject to } y_i(w^T \emptyset(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (4)$$

In the above equation, $C > 0$ is the penalty parameter of the error term, and ξ_i is a non-negative slack variable to allow mis-classification of difficult or noisy data points.

Hence, the hyper plane that optimally separates the data is the one that minimizes:

$$\Phi(w) = \frac{1}{2}\|w\|^2 \quad (5)$$

It is independent of b because, provided that equation (2) is satisfied (i.e. it is a separating hyper plane), changing b will move it in the normal direction to itself. Accordingly, the margin remains unchanged but the hyper plane is no longer optimal in that it would be nearer to one class than the other.

The solution to the optimization problem of equation (5) under the constraints of equation (2) is given by the saddle point of the Lagrange functional [25].

$$\Phi(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i \left(y^i[\langle w, x^i\rangle + b] - 1\right) \quad (6)$$

where w and α are the Lagrange multipliers. After some modifications and substitutions, the final solution is:

$$f(x) = sign(\langle w^*, x\rangle + b) \quad (7)$$

in which $w^* = \sum_{i=1}^{l} \alpha_i y_i x_i$

Here, if f (x) is positive, the new input data point **x** belongs to class 1 ($y_i = 1$), and if f (x) is negative, **x** belongs to class 2 ($y_i = -1$). However, in case of non-linear separation, the original input data is projected into a high dimensional feature space: ϕ: Rn→ Rd, n << d. i.e. $\mathbf{x} \rightarrow \phi(\mathbf{x})$, in which the input data can be linearly separated. In such a space, the dot-product from Eq. (7) is transformed into $\phi(\mathbf{x}i)$. $\phi(\mathbf{x})$, and the non-linear function can be expressed as [11]:

$$f(x) = sign(\sum_{i,j=1}^{N} \alpha_i y_i K(x_i, x_j) + b) \quad (8)$$

in which, $K(x_i, x_j) = (\varphi^T(x_i).\varphi(x_j))$ is the kernel function. In the case of a linear kernel, k is

the dot product. It may be useful to think of the kernel, $K(x_i, x_j)$, as comparing patterns or evaluating the proximity of objects in their feature space. Thus a test point is evaluated by comparing it with all training points [24]. Since choosing a suitable kernel for SVM comprises the building blocks of the machines, it is one of the most important steps in the SVMs model development [7].

Since the most common kernels stated in the literature that obtained the best improvements were RBF, polynomial, and linear, while other known kernels achieved surprisingly poor results [26], therefore, the following three basic kernels were utilized in this study:

- linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ (9)
- polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d, \gamma > 0$ (10)
- radial basis function(RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \gamma > 0$ (11)

In the above equations, d and γ are the kernel parameters. The optimum kernel function is generally selected through a trial and error procedure [27]. The original data matrix centered at their mean, and was scaled to have unit standard deviation before training.

2.5. Parameter optimization

There are two parameters for the radial basis function (RBF) kernel [28]: C (penalty parameter) and γ (a tuning parameter controlling the width of the kernel function). The correct choice of these parameters has a great influence on the stability and generalizing performance of the model.

It is not known beforehand which C and γ values are best for a given problem. Consequently, some kinds of model selection (parameter search) must be done. The goal is to identify good values for these parameters (C; γ) so that the classifier can accurately predict the unknown data (i.e. testing data). The γ value is important in the RBF model, leading to under- or over-fitting of the prediction model. An improved version of finding the optimum values for these parameters is through cross-validation [29, 30]. Since no assumption has been made about the data or noise distributions, cross-validation can be robust for tuning parameter selection [31]. The cross-validation procedure can prevent the over-fitting problem. In this research work, various pairs of values (C; γ) were tried, and the one with the best 5-fold cross-validation accuracy was picked. We used the simplex search method [32] for this purpose.

For the case of a polynomial kernel function, the only optimized parameter was the order of the equation. A good way of choosing the value ford (degree of the exponent in a polynomial kernel) is to start with 1 (a linear model), and increment it until the estimated error ceases to improve [13, 33]. This was implemented jointly with 5-fold cross-validation, and the value with the minimum out-of sample mis-classification error was selected as the optimum degree of the exponent value. The sequential minimal optimization (SMO) algorithm [34] was employed to train the SVMs model. In this algorithm, analytical solution of a subset can be obtained directly without invoking a quadratic optimizer, which is regarded as the main advantage of this method [12]. All of the required computations were done using MATLAB (R2013b).

2.6. K-nearest neighbor (K-NN)

The most basic and simplest instance-based method is the nearest neighbor (NN) inducer, which was first examined by Fix and Hodges [35]. It can be represented by the following rule: to classify an unknown pattern, choose the class of the nearest example in the training set as measured by a given distance metric. A common extension is to choose the most common class in the k-nearest neighbors (K-NNs) [18]. Its appeal stems from the fact that its decision surfaces are non-linear, there is only a single integer parameter (which is easily tuned with cross-validation), and the expected quality of predictions improves automatically as the amount of training data increases [36]. In the absence of a prior knowledge, most K-NN classifiers use simple euclidean distances to measure the dis-similarities between the examples represented as vector inputs. In this method, an unknown pattern is classified according to the majority of the class memberships of its K-nearest neighbors in the training set [37]. One of the most important optimization parameters in this modeling procedure is the number of nearest neighbors (k) because it should be the maximum number of neighbors with the minimum possible error [38]. This method has been utilized earlier in water quality researches [e.g. 6, 39, 40]. The mathematical detail of this method, though simple, is out of the scope of this work; however, interested readers can refer to the vast references existing in this field [e.g. 41].

2.7. Model evaluation

Having trained an optimized model, we applied it to the data in the model-testing data set to estimate several measures in order to evaluate the effectiveness of our method. These measures are classification accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. A confusion matrix [42] contains information about the actual and predicted classifications done by a classification system. Table 2 shows a typical confusion matrix for a two class classifier. Each one of the model evaluation criteria has been defined using elements of confusion matrix as:

$$\text{Positive predictive value} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Negative predictive value} = \frac{TN}{FN + TN} \quad (13)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (14)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (15)$$

$$\text{Accuracy}(\%) = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (16)$$

Table 2. A typical confusion matrix.

Test outcome		Condition	
		Condition positive	**Condition negative**
	Test outcome positive	True Positive	False positive(Type I error)
	Test outcome negative	False negative (Type II error)	True negative

2.8. Sensitivity analysis

The relative importance of each groundwater quality parameter in the classification of aquifer was considered using a sensitivity analysis. In the model building studies, sensitivity analysis generally refers to assessment of the importance of predictors in the fitted models. During this process, the variables are usually ranked according to the deterioration of the model performance criterion (e.g. specificity in this case) if a variable is removed from the model. This analysis is helpful for the identification of less important variables to be removed or ignored in the subsequent studies, in addition to the most essential variables [43]. The leave-one-out method was applied with SVMs, which corresponds to assess the changes in the error that would be obtained if each input variable was removed at a time. Each model was trained for ten times, and the average specificity was used for the sensitivity of each parameter.

3. Results and discussion

3.1. Cluster analysis

In this study, cluster analysis was used to categorize the groundwater quality sampling stations into appropriate clusters. The results obtained are expected to reveal the characteristics and extent of pollution for each cluster so that the spatial distribution of water pollution can be evaluated to implement the classification of the aquifer by other classification methods [44]. A dendogram, which clearly differentiates groups of objects, has small distances in the far branches of the tree, and large differences in the near branches.

Considering Figure 2 (dendrogram with average linkage method and Manhattan distance as similarity measure), it can be concluded that two groups can be identified. Cluster I (stations 1, 3, 7, 8, 10, 14, 19, 20, 22, 25, 26, 28, 37 and 40), which was called group A, and ClusterII (stations 2, 4, 5, 6, 9, 11, 12, 15, 16, 17, 18, 21, 23, 24, 27, 29, 30, 31, 32, 33, 34, 35, 36, 38, 39 and 41) that was called group B. In order to consider the difference between these two groups with respect to 17 groundwater quality variables, the Box-and-Whisker plots related to these stations were drawn. With respect to Figure 3a and Figure 3b, it can be concluded that, except for Cr and Mn, the amounts of other groundwater quality variables were higher in group B compared with those of group A. The different groups have been illustrated in Figure 3 (a, b).In addition, the location of sampling stations has also been given in Figure 4.

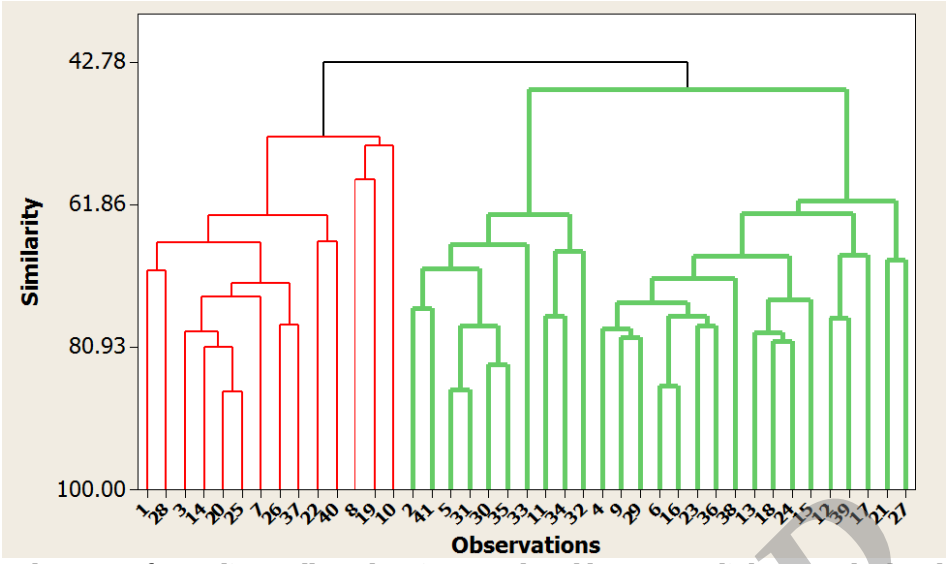

Figure 2. Dendrogram of sampling wells and springs produced by average linkage method and Manhattan distance measure.

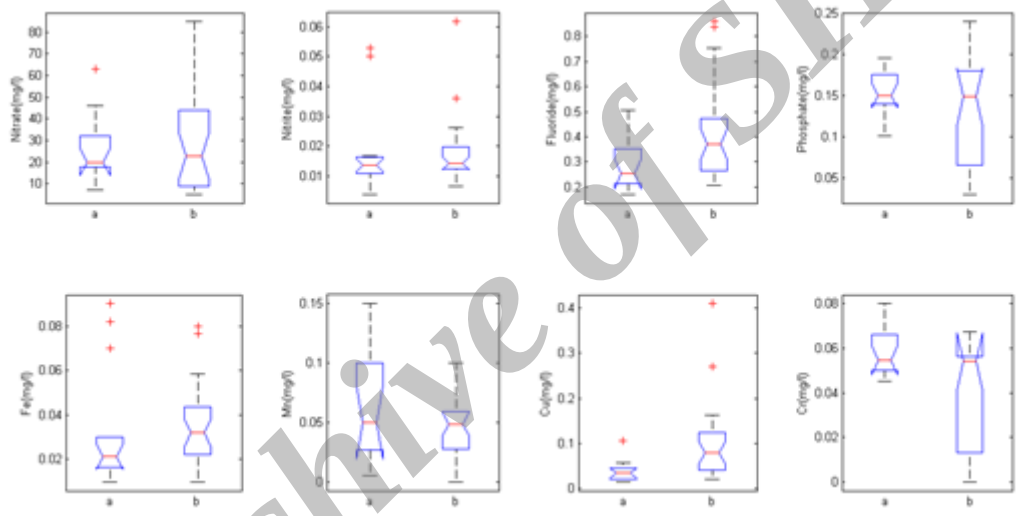

Figure 3. a). Box-and-Whisker plot of groundwater quality variables produced based on groups of cluster analysis(groundwater parameters are nitrate, nitrite, fluoride, phosphate, Fe, Mn, Cu and Cr).

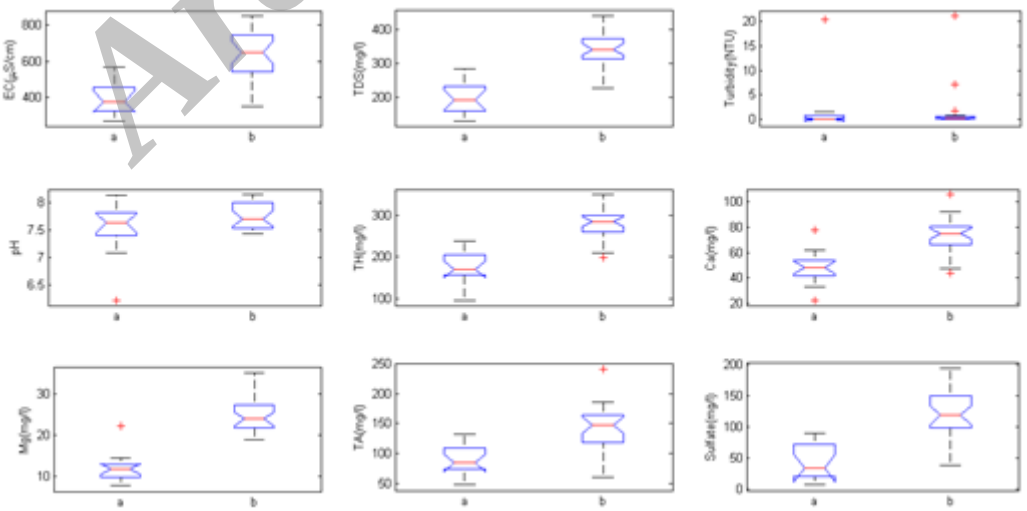

Figure 3. b). Box-and-Whisker plot of groundwater quality variables produced based on groups of cluster analysis(groundwater parameters are EC, TDS, turbidity, pH, TH, Ca, Mg, TA and sulfate).

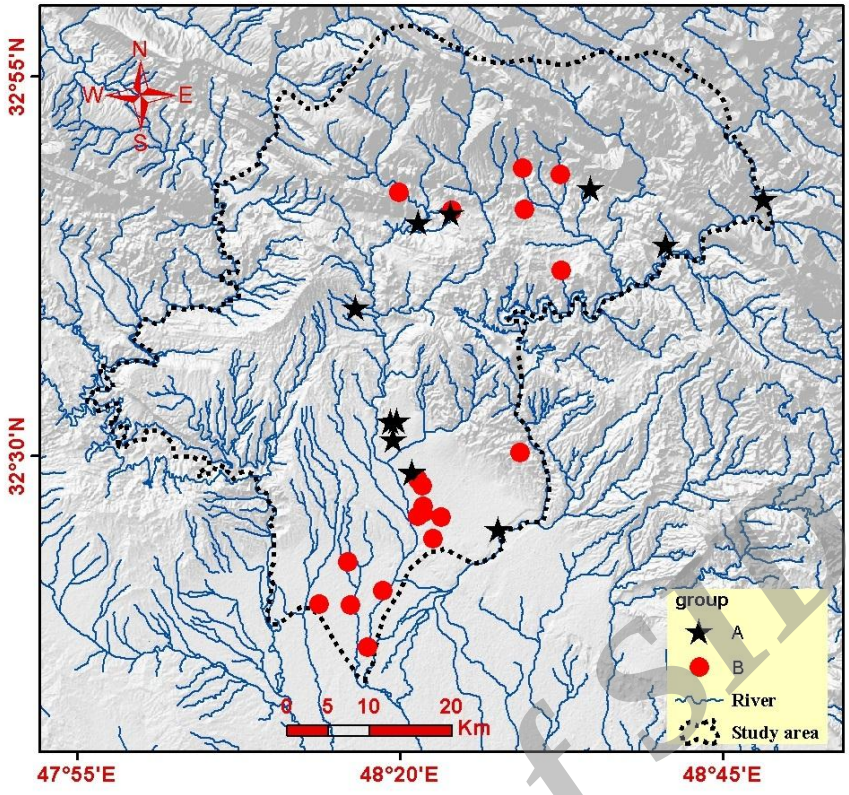

Figure 4. Difference in location of groundwater sampling stations for each group in studied area.

The water depth in the northern part of the studied area is higher than that in the southern part, ranging from 88m in the north western part of the area around Do-Koheh to about 3 m in the southern part around Haft-Tapeh [45]. The higher penetration of contamination to the shallow aquifer in the southern part may be one of the contributing factors for the concentration of group B in the southern part of the studied area. In addition, some of the stations in group A are located along the Dez River, which has an influence on the dilution of groundwater and reduction of the contamination level of these wells. In the next step, given the class for each groundwater station, K-NN and SVMs were applied for classification of the considered aquifer.

3.2. Factor analysis

Tke Kaiser-Meyer-Olkin (KMO) test is a representative test of the sampling adequacy to conduct factor analysis. There is no cut-off point associated with this test; however, if the test result is smaller than 0.5, the factor analysis is not suitable [46]. The result of KMO test was 0.703, indicating the suitability of factor analysis. In addition, chi-square distribution ($\chi 2$) of Bartlett's test of sphericity was high (647.59), and highly significant, implying the existence of a common factor among the relevant matrices of the parent population [46]. The results of factor analysis using principal component analysis and Varimax rotation method are given in Tables 3 and 4, respectively.

Table 3. Results of factor analysis.

Components	Eigenvalues	% of variance explained	Cumulative % of variance explained
1	5.59	32.88	32.88
2	3.25	19.13	52.01
3	1.24	7.31	59.32
4	1.22	7.16	66.48
5	1.17	6.86	73.34
6	1.08	6.35	79.69
7	1.01	5.94	85.63
8	0.91	5.34	90.97

Table 4. Rotated component matrix for each heavy metal in studied area.

Heavy metals	Component							
	1	2	3	4	5	6	7	8
EC	.956	-.079	.182	.031	-.052	.099	.013	.049
TDS	.929	.262	.080	-.032	-.077	.032	.013	.033
Turbidity	-.121	.196	-.028	-.067	-.007	-.953	-.069	-.006
pH	-.059	.336	-.587	-.417	.432	.231	-.091	-.079
TH*	.973	.103	.031	.009	-.008	.067	.040	.099
Ca	.900	-.070	.008	.070	.038	.056	.085	.283
Mg	.839	.333	.054	-.056	.140	.033	.040	-.066
TA**	.443	.754	.085	-.139	.041	-.059	-.142	-.020
Sulfate	.873	-.135	-.276	.007	-.112	-.040	-.016	.225
Nitrate	.165	-.436	.728	-.237	.153	.168	.090	.226
Nitrite	.093	-.196	.079	-.133	.079	.072	.950	.074
Flouride	.304	.686	-.378	.155	-.074	-.021	-.104	-.071
Phosphate	.144	-.846	.060	.223	.220	.224	.044	-.062
Fe	.045	.205	-.030	-.183	-.901	.010	-.089	-.074
Mn	.006	-.080	-.062	.910	.172	.084	-.136	-.010
Cu	.477	-.011	.191	-.008	.089	.002	.101	.825
Cr	.006	-.921	.216	-.007	.037	.039	.076	.031

*:TH stands for total hardness.
**:TH stands for total alkalinity.

There are many criteria for retaining the number of factors. For instance, according to Kaiser Criterion [47], only factors with eigenvalues greater than 1 are retained. However, Jolliffe [48] believed that Kaiser's criterion was too large, and suggested using a cut-off of 0.7 on the eigenvalues instead. Therefore, based on the Jolliffe's criterion, eight components were kept accounting for 90.97%of the total data variance.

The first factor encompasses 32.88%of the total variance, and has an eigenvalue of 5.59.Moreover, it is highly loaded with EC, TDS, total hardness, Ca, Mg and sulfate. In natural waters, EC depends mainly on the concentration of major ions such as Ca^{2+}, Mg^{2+}, Na^{+}, Cl^{-}, SO_4^{2-}, and HCO_3^{-} [49]. Moreover, electrical conductivity of water is a direct function of its total dissolved salts [50]. On the other hand, hardness is a property of cations (Ca^{2+} and Mg^{2+}), while alkalinity is a property of anions (HCO_3^{-} and CO_3^{2-}). The simultaneous high loading of EC, TDS, total hardness, Ca, Mg, and sulfate on the first factor might be due to the above-mentioned reasons. The first factor mainly implies the parameters that emanate from the geogenic sources in the aquifer. The second factor comprised 19.13%of the total variance, and has a highly positive loading with total alkalinity and fluoride, while a highly negative loading with phosphate and chromium. Chromium is one of the heavy metals that is included in phosphorus fertilizers as impurities, as has been proved by the results of the previous researches [e.g. 51-53].The high positive loading of this factor with phosphate and chromium may denote the role of agricultural activities on the contamination of groundwater due to the application of phosphorus fertilizers. The third factor accounting for 7.31%of the total variance has both a high negative loading with pH and a high positive loading with nitrate. It may indicate the role of pH on the nitrification of nitrogenous forms in the aquifer. Nitrate (NO_3^{-}) is one of the several inorganic pollutants contributed by the nitrogenous fertilizers, organic manures, human and animal wastes, and industrial effluents through the biochemical activities of microorganisms. Excessive use of nitrogenous fertilizers in agriculture has been one of the primary sources of high nitrate in groundwater [54]. Nitrification is relatively sensitive to pH, in part, because of the generation of ammonia (NH_3) under alkaline conditions and nitrous acid (HNO_2) under acidic conditions [55]. It is reasonable to infer that pH 6.58 is the optimum pH range for nitrification but rates are likely to significantly decreased below pH 6.0 or above pH 8.5 [55]. The fourth and fifth factors have a high positive and negative loading with manganese and iron variables. The main sources of iron in ground water are natural as a mineral from sediment and rocks or from mining, industrial waste, and corroding metal [56]. To the contrary, natural sources of manganese are more common in deeper

wells where the water has been in contact with rock for a longer time. In these anaerobic conditions, manganese is released from minerals, and reduced to its more soluble form, Mn (II). This form is apparently the most soluble one in most waters [57]. The sixth and seventh factors encompass high negative and positive loadings with turbidity and nitrite, respectively. Finally, copper was highly loaded with the eighth factor. The last factor just accounts for 5.34%of the total variance, implying the minor importance of this factor in comparison with that of the other factors. These eight factors accumulatively account for 90.97%of the total variance, thus the variationsin17 variables can be covered by just eight factors.

3.3. Support vector machines (SVMs)

Referring to SVMs, as stated earlier, there were two parameters to be optimized for the radial basis function (RBF), kernel (C and γ), and one parameter for polynomial kernel (d). For the RBF method, the optimum values for penalty parameter (C) and gamma (γ) that yielded the minimum value of kernel function were 10.49 and 0.54, respectively. On the other hand, for the case of polynomial kernel, a trial and error procedure was followed to obtain the optimum polynomial exponent. The results obtained are given in Table 5.

Table 5. Performance of SVMs for polynomial kernel function with respect to degree of exponent.

Degree of exponent	Performance				
	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy
1	1.00	0.87	0.92	1.00	0.94
2	1.00	0.64	0.83	1.00	0.86
3	0.99	0.66	0.88	0.90	0.90
4	0.92	0.55	0.85	0.72	0.80
5	1.00	0.55	0.74	1.00	0.82

These results indicated that the best performed model was related to an exponent of degree one with an accuracy of 94% for the test data set, in which the sensitivity and specificity were 1.00 and 0.87, respectively. This shows that linear SVMs has out-performed compared with that of non-linear models. Using other exponents of the polynomial kernels (non-linear kernels) increased the risk of over-fitting, leading to a poor out-of-sample prediction. Moreover, as it can be seen in Table 5, there is no significant difference among the results of different kernels, indicating that an acceptable result can be achieved by selecting optimum parameters of a kernel. The same results were obtained by Sadeghi et al. [13] in their work on the use of SVMs to predict distribution of an invasive water fern in Anzali wetland, Iran.

In addition, another decision that had to be made was the subdivision of the data set into different subsets, which were used for training and testing. In order to examine the generalization ability of a model, a separate data set had to be used for training and testing the model [58]. The best option was to divide the data randomly into two parts. This procedure of partitioning data is sometimes named the hold-out procedure [59]. Different data divisions were tried in this study to find the optimum method based on the out-of-sample generalization error of the model, and the performances for ten times retraining of each model are rendered in Table 6 [43].

Table 6. Results of different division of original data set for SVMs and K-NN methods.

Percent of training data	Training method	Performance				
		Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy
75%	SVMs	1.00	0.97	0.98	1.00	0.99
65%		0.98	0.94	0.97	0.95	0.95
55%		0.98	0.85	0.93	0.95	0.93
75%	K-NN	0.97	0.90	0.94	0.94	0.94
65%		0.91	0.87	0.94	0.77	0.89
55%		0.87	0.77	0.88	0.76	0.83

According to this table, the best data division was75% for the training part, and 25% for the testing part, leading to an accuracy of 99%. Moreover, this data division resulted in sensitivity (rate of correctly classified objects for each class) and specificity (rate of correctly rejected objects for each class) of 1.00 and 0.97 for SVMs, respectively. The same results were obtained for the K-NN method, resulting in 94% accuracy in predictions besides sensitivity and specificity of in turn 0.97 and 0.90. The overall results of SVMs for each kernel method using optimum parameter values and ten-times retraining of each model are given in Table 7.

The performance of this table is based upon the test data set. The results show that the performance of the three models is nearly the same. In view of sensitivity, for instance, the polynomial kernel yielded the highest value, indicating perfect classification of sampling stations. The sensitivity in this case is high because the true negative result in this prediction is high as it does not wrongly predict the data that was not supposed to be predicted in a wrong class; however, the accuracy shows that the prediction is roughly accurate for all the classes. In studies carried out by other researchers such as Najah et al. [60] and Liu et al. [61] using SVMs for the classification of water quality data, small values of error were produced, and an accuracy above 70% was obtained. RBF kernel non-linearly maps samples into a higher dimensional space, thus unlike the linear kernel, it can handle the case when the relation between the class labels and attributes is non-linear [15]. However, as concluded in the above-mentioned section, linear SVMs have been able to discriminate between the classes efficiently, and so the performances of these two models havenot been that much different.

Table 7. Comparison of results of SVMs and K-NN for ten-times retraining of each model.

Model types	Model evaluation				
	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy
SVMs(RBF)	0.97	0.91	0.96	0.93	0.95
SVMs(Polynomial)	1	0.87	0.92	1	0.94
SVMs(Linear)	0.93	0.97	0.92	0.89	0.95
K-NN	0.90	0.88	0.93	0.88	0.93

3.4. K-NN classifier

One of the most important parameters to optimize for the K-NN method is the tuning of the nearest neighbors involved in the k-nearest neighbor classifiers, which clearly constrains over-fitting. Both lower and higher than enough number of nearest neighbors may contribute to over-fitting and under-fitting, respectively [62]. The problem of over-fitting in classification problems is that perfect training performance by no means predicts the same performance of the trained classifier on unseen objects [62]. In addition, the basic assumption underlying over-fitting prevention schemes is that simpler classification models are better than the more complex ones(especially in situations where the errors on the training data are equal). The results of this study showed that the best mis-classification error was for five nearest neighbors, which mis-classified 7%of the test dataset (Figure 5).

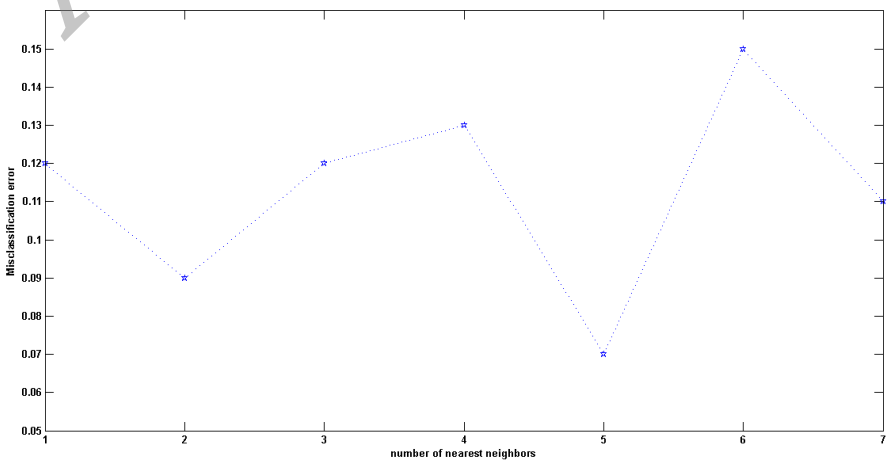

Figure 5. Number of nearest neighbors versus amount of mis-classification error for K-NN method.

Given this optimum value, the data was trained by the k-NN classifier for which the results are rendered in Table 7.These results showed a lower efficiency compared with those for SVMs, where sensitivity and specificity reduced to 0.90 and 0.88, respectively, although the accuracy of this model was93%. The same result has beenobtained by Modaresi and Araghinejad [10] in their study for water quality classification using K-NN, SVMs, and probabilistic neural network (PNN), in which the best results were yielded by SVMs followed by PNN, and the worst performing model was K-NN.

3.5. Comparison of methods

As it is obvious from the results of this study, and also confirmed by Khalil et al. [7], the SVM model is characterized by a highly effective mechanism for avoiding over-fitting that results in a good generalization. It can be a plague when working with a small number of data records like the case of this study. However, generalization of this method proved the feasibility of this modeling procedure for working with such data set [13]. This is especially important as the high cost of sampling and analysis of water quality parameters is an obstacle for gathering a large water quality data set, especially in developing countries. As mentioned by Vapnik [14],one of the reasons for better generalization of SVMs compared with that of other classifiers is that SVMs simultaneously minimizes the empirical classification error, and maximize the geometric margin. Thus the larger the margin, the better the generalization error of the classifier would be [14].

To consider the importance of water quality variables in classification of the studied aquifer, sensitivity analysis was implemented. In this field, the variables that change the output more when tweaked are more sensitive, and, therefore, more important. On the other hand, the features for which the predictions do not vary a lot when they are tweaked are considered less important, and can be disregarded for the following monitoring programs or pruned out during modeling as a method of feature selection.

3.6. Sensitivity analysis

The results of sensitivity analysis are illustrated in Figure 6. With respect to this figure, it can be concluded that calcium next to nitrate are the most influential parameters in the classification of groundwater with SVMs using specificity as mis-classification criterion. In general, agricultural lands cover over 80% of the studied area, and this aquifer is dominantly an unconfined aquifer [63]. Nitrate concentration is the greatest in these areas, especially if they are heavily irrigated, as most parts of this aquifer are well-drained. In the recent years, the application of fertilizers and manure in agricultural fields of the studied area has significantly increased, resulting in the elevated levels of nitrate in groundwater resources [64]. In addition, the low correlation coefficient between nitrate and calcium revokes the hypothesis that they have originated from the same source (e.g. agricultural fields). Referring to geological formations in the studied area (Figure 7), it can be concluded that different kinds of limestone including Gachsaran formation, which mainly consists of argillaceous limestone and limestone, are prevalent in the studied area. These geological formations are most likely the main contributing factors for the concentrations of calcium in groundwater. Since the dominant groundwater flow direction is from the north to the south, probably, the upper area has a significant influence on the lower parts of the aquifer.

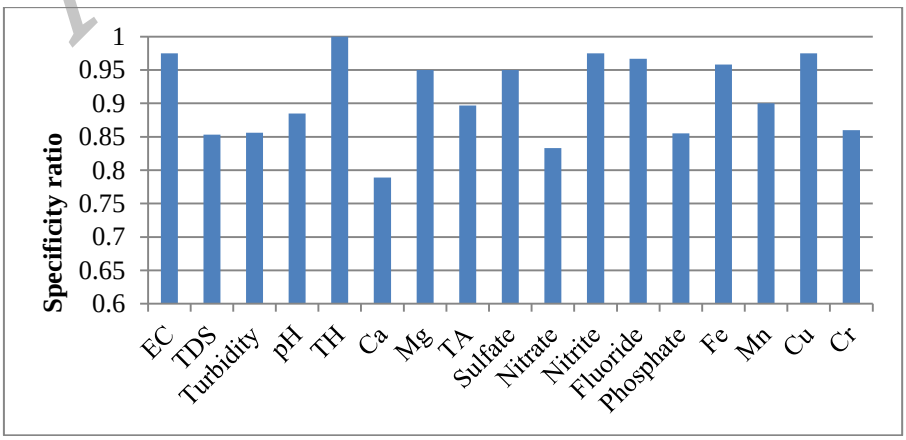

Figure 6. Sensitivity of 17 groundwater quality variables based on specificity using SVMs as base classifier.

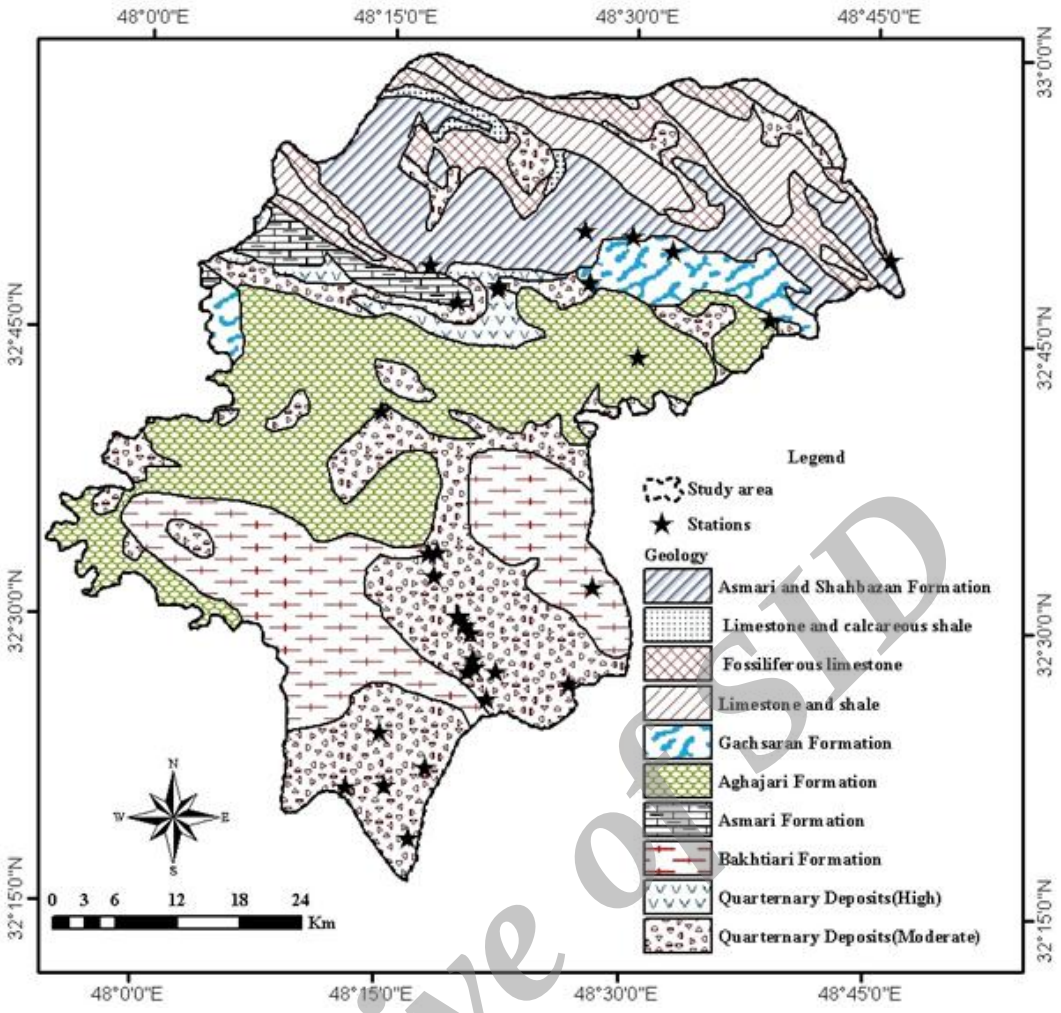

Figure 7. Dominant geological formations in studied area.

4. Conclusions

Groundwater quality classification is a tool for local managers to use in land-use management decisions. Among these methods, support vector machines (SVMs) is one the most recently applied classification methods in environmental researches. In this study, the performance of SVMs for classification of a groundwater quality data set (2006-2013) in Andimesk Aquifer was compared with that of K-NN. As a whole, SVMs proved to have both a better performance and better generalization ability, especially for small data set.

Acknowledgments

The authors are grateful to the helps of Andimeshk Health Network and Iran Ministry of Energy for providing us with water quality data.

References

[1]. Motagh, M.T.R., Walter, M.A., Sharifi, E., Fielding, A., Schenk, J. and Anderssohn Zschau, J. (2008). Land subsidence in Iran caused by widespread water reservoir overexploitation, Geophys. Res. Lett.35:L16403,doi:10.1029/2008GL033814.

[2]. Voss, K.A., Famiglietti, J.S., Lo, M., de Linage, C., Rodell, M. and Swenson, S.C. (2013). Groundwater depletion in the Middle East from GRACE with implications for transboundary water management in the Tigris-Euphrates-Western Iran region, Water Resour. Res. 49: doi:10.1002/wrcr.20078.

[3]. Baghvand, A., Nasrabadi, T., Nabi Bidhendi, G., Vosoogh, A., Karbassi, A. and Mehrdadi, N. (2010). Groundwater quality degradation of an aquifer in Iran central desert, Desalination. 260: 264-275.

[4]. Nasrabadi, T. and Abbasi Maedeh, P. (2014). Groundwater quality degradation of urban areas (case study: Tehran city, Iran. Int. J. Environ. Sci. Tech. 11: 293-302.

[5]. Wallace, J., Lowe, M., King, J.K., Sabbah, W. and Thomas, K.J. (2012). Hydrogeology of Morgan Valley, Morgan County, Utah, Utah Geological Survey.

[6]. Vallejuelo, S.F.O.D., Arana, G., Diego, A.D. and Madariaga, J.M. (2011). Pattern recognition and classification of sediments according to their metal

content using chemometric tools. A case study: The estuary of Nerbioi-Ibaizabal River (Bilbao, Basque Country), Chemosphere, 85: 1347-1352.

[7]. Khalil, A., Almasri, M.N., McKee, M. and Kaluarachchi, J.J. (2005). Applicability of statistical learning algorithms in groundwater quality modeling. Water Resour Res 41, W05010, doi:10.1029/2004WR003608.

[8]. Singh, K.P., Malik, A., Singh, V.K., Mohan, D. and Sinha, S. (2005). Chemometric analysis of groundwater quality data of alluvial aquifer of Gangetic plain, North India, Analytica Chimica. Acta. 550: 82-91.

[9]. Spruill, T.B., Showers, W.J. and Howe, S.S. (2002). Application of classification-tree methods to identify nitrate sources in ground water. J. Environ. Qual. 31: 1538-1549.

[10]. Modaresi, F. and 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: 4095-4111.

[11]. Singh, K.P., Basant, N. and Gupta, S. (2011). Support vector machines in water quality management. Analytica. Chimica. Acta. 703: 152-162.

[12]. Yoon, H., Jun, S.H., Hyun, Y., Bae, G.O. and Lee, K.K. (2011). A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, J. Hydrol. 396: 128-138.

[13]. Sadeghi, R., Zarkami, R., Sabetraftar, K. and Van Damme, P. (2012). Use of support vector machines (SVMs) to predict distribution of an invasive water fern Azolla filiculoides (Lam.) in Anzali wetland, southern Caspian Sea. Iran, Ecol. Model. 244: 117-126.

[14]. Vapnik, V. (1999). The Nature of Statistical Learning Theory, Second Edition, Springer, New York.

[15]. Hsu, C.W., Chang, C.C. and Lin, C.J. (2010). A practical guide to support vector classification.Tech. rep., Department of Computer Science, National Taiwan University.

[16]. Guo, Q., Kellya, M. and Graham, C.H. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. Ecol. Model. 182: 75-90.

[17]. Akay, M.F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis, Expert Syst. Appl. 36: 3240-3247.

[18]. Rokach, L. (2010). Pattern classification using ensemble methods, World Scientific Publishing, 225 P.

[19]. Rezania, A.R., Naseri, A.A. and Albaji, M. (2009). Assessment of soil properties for irrigation methods in North Andimeshk Plain, Iran. J. Food Agric. Environ. 7: 728-733.

[20]. Nouri, J., Mahvi, A.H., Babaei, A.A., Jahed, G.R. and Ahmadpour, E. (2006). Investigation of heavy metals in groundwater. P. J. Bio. Sci. 9: 377-384.

[21]. Panda, C.P., Sundaray, S.K., Rath, P., Nayak, U.C. and Bhatta, D. (2006). Application of Factor and Cluster Analysis for Characterization of River and Estuarine Water Systems – a Case Study: Mahanadi River (India). J. Hydrol. 331: 434-445.

[22]. Akbar T.A., Hassan, Q.K. and Achari, G. (2011). A Methodology for Clustering Lakes in Alberta on the basis of Water Quality Parameters. Clean – Soil, Air, Water. 39 (10): 916-924.

[23]. Araújo, R., Nunes, U., Oliveira, L., Sousa, P. and Peixoto, P. (2008). Support Vector Machines and Features for Environment Perception in Mobile Robotics, Coimbra, Portugal.

[24]. Xuan, W., Jiake, L. and Deti, X. (2010). A Hybrid Approach of Support Vector Machine with Particle Swarm Optimization for Water Quality Prediction, The 5th International Conference on Computer Science & Education, China.

[25]. Minoux, M. (1986). Mathematical Programming: Theory and Algorithms. John Wiley and Sons.

[26]. Gaspar, P., Carbonell, J. and Oliveira, J.L. (2012). On the parameter optimization of Support Vector Machines for binary classification. J Integr Bioinform. 9 (3): 1-11.

[27]. Widodo, A. and Yang, B.S. (2007). Support vector machine in machine condition monitoring and fault diagnosis, Mech. Syst. Sig. Pro. 21(6): 2560-2574.

[28]. Li, H., Liang, Y. and Xu, Q. (2009). Support vector machines and its applications in chemistry, Chem. Intell. Lab. Syst. 95:188-198.

[29]. Browne, M.W. (2000). Cross-validation methods. J. Math.Psychol. 44:108-132.

[30]. Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection, Statistics Surveys. 4: 40-79.

[31]. Atkenson, C.G., Moore, A.W. and Schaal, S. (1997). Locally weighted learning, Artif. Intel. Rev. 11: 11-73.

[32]. Lagarias, J.C.. Reeds, J.A.. Wright, M.H. and Wright, P.E. (1998). Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," SIAM Journal of Optimization9. (1): 112-147.

[33]. Hoang, H., Lock, K., Mouton, A. and Goethals, P.L.M. (2010). Application of classification trees and support vector machines to model the presence

of macroinvertebrates in rivers inVietnam. Ecol. Inform. 5: 140-146.

[34]. Platt, J.C. (1999). Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smolar, A.J. (Eds.), Advances in Kernel Methods-Support Vector Learning. MIT Press, Cambridge, Massachusetts, USA.

[35]. Fix, E. and Hodges, J.L. (1957). Discriminatory analysis. Nonparametric discrimination.Consistency properties. Technical Report 4, US Air Force School of Aviation Medicine. Randolph Field, TX.

[36]. Goldberger, J., Roweis, S., Hinton, G. and Salakhutdinov, R. (2005). Neighbourhood components analysis.In L. K. Saul, Y. Weiss, and L. Bottou, editors. Cambridge, MA, MIT Press, Advances in Neural Information Processing Systems. 17: 513-520.

[37]. Zheng, W. and Tropsha, A. (2000). Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle.J. Chem. Inf. Comput. Sci. 40: 185-194.

[38]. Ruiz-Jimenez, J., Priego-Capote, F., Garcia-Olmo, J. and Luque de Castro, M.D. (2004). Use of chemometrics and mid infrared spectroscopy for the selection of extraction alternatives to reference analytical methods for total fat isolation. Analytica. Chimica. Acta. 525: 159-169.

[39]. Librando, V. (1991).Chemometric evaluation of surface water quality at regional level, Fresen. J. Anal. Chem. 339: 613-619.

[40]. Lee, B.H. and Scholz, M. (2006). A comparative study: Prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks, Water Air Soil Poll. 174 (1-4): 279-301.

[41]. Okun, O. (2011). Feature selection and ensemble methods for bioinformatics, algorithmic classification and implementations, Med. Info. Sci. Ref. 445 P.

[42]. Kohavi, R. and Provost, F. (1998). Glossary of terms. Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. 30 (2-3).

[43]. Gazzaz, N.M., Yusoff, M.K., Aris, A.Z., Juahir, H. and Ramli, M.F. (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors, Mar Pollut Bull. 64: 2409-2420.

[44]. Boyacioglu, H. and Boyacioglu, H. (2008). Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin. Turkey, Environ Geol. 54: 275-282.

[45]. Khodaei, K., Mohamadzadeh, H., Naseri, H.R. and Shahsavari, A.R. (2012). An Investigation on nitrate pollution in Dezful-Andimeshk plain and pollution sourcing using 14N and 18O radioisotopes. Iran Journal of Geology. 27: 93-111 (in Persian).

[46]. Wu, E.M.Y. and Kuo, S.L. (2012). Applying a multivariate statistical analysis model to evaluate the water quality of a watershed. Water Environ. Res. 84: 2075-2085.

[47]. Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement. 20: 141-151.

[48]. Jolliffe, I.T. (1972). Discarding variables in principal component analysis. I: Artificial data. J Appl Stat. 21: 160-173.

[49]. Tizro, A.T. and Voudouris, K.S. (2008). Groundwater quality in the semi-arid region of the Chahardouly basin, West Iran, Hydrol. Process. 22: 3066-3078.

[50]. Harilal, C.C., Hashim, A., Arun, P.R. and Baji, S.J. (2004). Ecology Environ Conservation. 10 (2): 187-192.

[51]. Watanabe, H. (1984). Accumulation of chromium from fertilizers in cultivated soils. Soil Sci Plant Nutr. 30 (4): 543-55.

[52]. Ciavatta, C. and Sequi, P. (1989). Evaluation of chromium release during the decomposition of leather meal fertilizers applied to the soil. Nutr Cycl Agroecosys. 19 (1): 7-11.

[53]. Carnelo, L.G.L., Miguez, S.R. and Marban, L. (1997). Heavy metals input with phosphate fertilizers used in Argentina.Sci Total Environ. 204 (3): 245-250.

[54]. Majumdar, D. and Gupta, N. (2000). Nitrate pollution of groundwater and associated human health disorders. Indian Journal of Environmental Health. 42 (1): 28-39.

[55]. USEPA. (1993). Manual. Nitrogen control. Report EPA/625/-93/010. US Environmental Protection Agency, Washington DC.

[56]. Roger, M. (1982). Ground water and the rural homeowner, Pamphlet, U.S. Geological Survey.

[57]. Nadaska, G., Lesny, J. and Michalik, I. (2012). Environmental Aspect of Manganese Chemistry. available at: http://heja.szif.hu/ENV/ENV-100702-A/env100702a.pdf.

[58]. Lek, S. and Guegan, J.F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120: 65-73.

[59]. Kohavi, R. and Wolpert, D.H. (1996). Bias plus variance decomposition for zero-one loss functions. In: Saitta, L. (Ed.),Machine learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, Bari, Italy. 275-283.

[60]. Najah, A., El-Shafie, A., Karim, O.A. and Jaafar, O. (2011). Integrated versus isolated scenario for prediction dissolved oxygen at progression of water

quality monitoring stations, Hydrol. Earth Syst. Sci. 15: 2693-2708.

[61]. Liu, J.P., Chang, M.Q. and Ma, X.Y. (2009). Groundwater Quality Assessment Based on Support Vector Machine. HAIHE River Basin Research and Planning Approach-Proceedings of 2009 International Symposium of HAIHE Basin Integrated Water and Environment Management, Beijing, China. 173-178.

[62]. Veenman, C.J. and Reinders, M.J.T. (2005). The nearest sub-class classifier: a compromise between the nearest mean and nearest neighbor classifier, IEE T. Pattern Anal. 27 (9): 1417-1429.

[63]. Shahsavari, A.A., Khodaei, K., Hatefi, R., Asadian, F. and Zamanzadeh, S.M. (2014). Distribution of total petroleum hydrocarbons in Dezful aquifer, Southwest of Iran. Arab. J. Geosci. 7: 2367-2375.

[64]. Mahvi, A.H., Nouri, J., Babaei, A.A. and Nabizadeh, R. (2005). Agricultural activities impact on groundwater nitrate pollution, Int. J. Environ. Sci. Tech. 2: 41-47.

مطالعه مقایسه‌ای بین کارایی روش‌های ماشین بردار پشتیبانی و الگوریتم K نزدیک‌ترین همسایه برای کلاسه‌بندی آب زیرزمینی

محمد ساکی زاده[1*] و روح‌الله میرزایی[2]

۱- گروه بهداشت و محیط‌زیست، دانشکده علوم، دانشگاه تربیت دبیر شهید رجایی، ایران

۲- گروه محیط‌زیست، دانشگاه کاشان، ایران

ارسال ۲۰۱۵/۶/۱۷، پذیرش ۲۰۱۵/۸/۳۰

* نویسنده مسئول مکاتبات: msakizadeh@gmail.com

چکیده:

هدف از این تحقیق، بررسی کارایی روش‌های ماشین بردار پشتیبانی و الگوریتم K نزدیک‌ترین همسایه برای کلاسه‌بندی یک آبخوان در استان خوزستان است. برای این منظور، ۱۷ پارامتر کیفیت آب زیرزمینی شامل هدایت الکتریکی، کل جامدات محلول، کدورت، pH، سختی کل، کلسیم، منیزیم، قلیایت کل، سولفات، نیترات، نیتریت، فلوراید، فسفات، آهن، منگنز، مس و کروم در ۴۱ چاه و چشمه نمونه‌برداری شده در طول یک دوره ۸ ساله بین سال‌های ۱۳۸۵ تا ۱۳۹۲ مورد استفاده قرار گرفت. تحلیل خوشه‌ای منجر به تولید درختواره‌ای شد که ایستگاه‌های موجود را به دو گروه کلی تقسیم‌بندی کرد. تحلیل عامل منجر به استخراج ۸ عامل شد که در مجموع دربرگیرنده ۹۰.۹۷ درصد از واریانس داده‌های اولیه شد. بر این اساس، تغییرات موجود بین ۱۷ پارامتر کیفی را می‌توان تنها با ۸ عامل پوشش داد. روش‌های ماشین بردار پشتیبانی و الگوریتم K نزدیک‌ترین همسایه برای کلاسه‌بندی آبخوان تحت مطالعه مورد استفاده قرار گرفتند. نتایج روش ماشین بردار پشتیبانی نشان داد که بهترین مدل از نوع نمایی درجه یک بوده که از کارایی ۹۴ درصد برای داده‌های آزمون و از حساسیت و وضوح به ترتیب ۱ و ۰.۸۷ برخوردار بوده است. همچنین تفاوت معنی‌داری بین نتایج مدل‌های مختلف با تعداد متفاوت کرنل مشاهده نشده است که نشان می‌دهد مدل قابل قبول را می‌توان با انتخاب پارامترهای بهینه جهت کرنل به دست آورد. نتایج الگوریتم K نزدیک‌ترین همسایه از کارایی کمتری نسبت به روش‌های ماشین بردار پشتیبانی برخوردار بودند که میزان حساسیت و وضوح به ترتیب به ۰.۹۰ و ۰.۸۸ کاهش یافتند هرچند که میزان کارایی مدل برابر با ۹۳ درصد بود. آنالیز حساسیت بر روی پارامترهای کیفیت آب زیرزمینی انجام گرفت که نشان داد کلسیم در کنار نیترات مهم‌ترین پارامترهای تأثیرگذار بر روی طبقه‌بندی آبخوان به شمار می‌روند.

کلمات کلیدی: آب زیرزمینی، ماشین بردار پشتیبانی، الگوریتم K نزدیک‌ترین همسایه، توابع کرنل.