Neural Comput & Applic (2017) 28:3229–3238
DOI 10.1007/s00521-016-2231-x

ORIGINAL ARTICLE

Support vector machine and artificial neural network to model soil pollution: a case study in Semnan Province, Iran

Mohamad Sakizadeh[1] · Rouhollah Mirzaei[2] · Hadi Ghorbani[3]

Received: 24 February 2015 / Accepted: 16 February 2016 / Published online: 3 March 2016

Abstract To study the extent of soil pollution in Shahrood and Damghan located in Semnan Province, Iran, 229 soil samples were taken and the levels of 12 heavy metals (Ag, Co, Pb, Tl, Be, Ni, Cd, Ba, Cu, V, Zn and Cr) were analyzed. Elevated values of some heavy metals such as Cr, Ni and V were detected in the study area. In order to predict soil pollution index (SPI) with respect to the concentration levels of 12 detected heavy metals, support vector machines (SVMs) with different kernels (linear, RBF and polynomial) and artificial neural networks (ANNs) were utilized. The database was repeatedly randomly split into training and testing data sets, and both SVMs and ANNs were trained and tested for each split. The testing results of the support vector regression (SVR) model with combinations of parameter sets were compared to optimize the parameters of SVMs with different kernels. The out-of-sample generalization ability of different kernels was roughly high and the same. Therefore, RBF kernel was selected for comparison with ANNs with early stopping. The correlation coefficients between the predicted and observed SPI for the RBF kernel and ANN with early stopping were 0.997 and 0.995, implying the same performance of these two methods. The results indicated that because of some problems associated with ANNs (such as local minima), for cases in which there are quite comparable results for ANNs and SVMs, the usage of SVMs is preferable.

Keywords Heavy metals · Soil pollution index · Support vector machines · Artificial neural network

✉ Mohamad Sakizadeh
msakizadeh@gmail.com

[1] Department of Environmental Sciences, Faculty of Sciences, Shahid Rajaee Teacher Training University, Shahid Shabanloo Avenue, Lavizan, Tehran 1678815811, Iran

[2] Department of Environmental Sciences, University of Kashan, Kashan, Iran

[3] University of Shahrood, Shahrood, Iran

1 Introduction

All soils naturally contain trace levels of metals. The presence of metals in soil is, therefore, not indicative of contamination. However, if the level of heavy metals exceeds some ranges, it might be considered as a potential risk for the human health. The soil heavy metals in the environment are relatively stable and are difficult to remove through natural processes [1]. Thus, the monitoring and assessment of the environmental quality of soils play an important role in restoring damaged ecosystems and protecting soil environmental quality. Many calculation methods have been presented to assess the environmental quality of soil, such as geo-accumulation index [2], principle component analysis [3], integrated pollution index [4] and maturity index [5]. In addition, up until now, there have been some studies on the concentration of heavy metals in soil samples in different parts of Iran [6–9]. On the contrary, several calculations have been made using artificial neural networks (ANN) [10–18] and support vector machines (SVMs) [19–22] in different fields of environmental sciences. Moreover, considering the researches in the field of soil science, recently some of the most important soil properties such as cation exchange capacity (CEC), field capacity (FC) and permanent wilting point (PWP) which are hard to measure in the field are being predicted through more readily available soil properties such as particle-size distribution (sand, silt and clay

content), organic matter or organic *C* content, bulk density, porosity, etc. by neural network [23] and SVMs [24] methods via Pedotransfer Functions. Some of the other applications of ANN and SVM in soil management include prediction of soil hydraulic conductivity [25], soil moisture [26, 27] and soil organic carbon [28].

Support vector machine, based on the structural risk minimization (SRM) principle, seems to be a promising method for data mining and has been used for both classification and regression problems. The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

On the other hand, artificial neural network is a type of artificial intelligence (computer system) that attempts to mimic the way the human brain processes and stores information. The ANNs are considered as standard nonlinear estimators, and their abilities have been verified in a variety of fields [29]. Details of the neural network, including different algorithms for network training, can be found in the extensive published literatures in this field [30–32].

The main difference between these two artificial intelligence methods is due to the algorithm used for reducing the generalization error. In SVM, the empirical risk minimization (ERM) in ANN is replaced by the SRM principle, which seeks to minimize an upper bound of the generalization error rather than minimize the training error [33]. The advantages of artificial neural network over traditional statistical techniques as explained by Peng and Wen [34] include: (1) Neural network is more accurate than statistical techniques especially when dealing with incomplete data records. (2) As the neural network can develop its own weighting scheme, so it is faster than other statistical techniques. (3) There is no need for prior knowledge during ANN's modeling, so it is a more flexible and powerful tool, while besides the main redeeming features of artificial neural network, the advantage of the SVM is the elimination of the local minimum issue of ANN.

To the best of our knowledge, SVM has not been used for the prediction of soil quality in the previous studies. In this research, two objectives are followed (a) to study the extent of soil pollution in Shahrood and Damghan located in Semnan Province, Iran, and (b) to predict the soil pollution index (SPI) by SVM and ANN and compare the performance of these two methods, accordingly.

2 Materials and methods

The study site covers an area of about 65,760 km^2 and a population of about 325,000 inhabitants resides in the study area. It is located in Semnan Province, in central part of Iran. Shahrood and Damghan are the main cities in this region. The area is characterized by a mountainous climate, having an average precipitation of 133.7 mm per year and the temperature varies from about –10 °C in winter to about 40 °C in summer [35].

One of the dominant activities in this region is mining and coal washing operations (e.g., Tazareh coal mine area is located 31 km far from Damghan City) with a high capability for soil pollution by heavy metals. To study the soil pollution, a systematic random sampling approach [36] was followed and a total of 229 soil samples were taken from the top 10 cm of the soil. Samples were collected in clean, dry plastic containers and were protected from contamination until preparation. Following dryness of samples in 60 °C, they were sieved through a <2-mm stainless steel mesh. The soil samples were then digested with nitric acid (HNO_3) and hydrochloric acid (HCl) in a ratio of 3:1 (HNO_3:HCl). Finally, Ag, Co, Pb, Tl, Be, Ni, Cd, Ba, Cu, V, Zn and Cr were analyzed by inductively coupled plasma (ICP) optical emission spectroscopy (ICP-OES).

In order to show the relative magnitudes of soil pollution, Nemerow's synthetical pollution index (P_n) for all the soil sampling points was calculated. Higher value for P_n indicates more serious pollution. The index has been utilized in the previous studies [37], and in the present research, using Iranian Soil standards of the Department of Environment (DOE) (Table 1), the index was calculated and applied as the soil quality assessment criterion. This index is calculated with the following equations:

$$P_n = \sqrt{(\max P_i^2 + \text{average}\, P_i^2)/2} \quad (1)$$

where

$$P_i = \frac{C_i}{S_i} \quad (2)$$

In equations, P_n is the Nemerow's synthetical pollution index, P_i is the pollution index of the *i*th heavy metal, C_i and S_i are the measured and assigned standard of the ith heavy metal, whereas max P_i and average P_i are the maximum and average values of pollution indices for all considered heavy metals, respectively [37].

Support vector regression (SVR) was one of the modeling procedures used in this study to predict the SPI given 12 heavy metals as the features. A popular regression version of SVM, ε-SVM, is used to find a function that has at most ε deviations from the actual obtained targets for all the training data, and is as flat as possible.

For building SVR forecasting model, the LIBSVM package proposed by Chang and Lin [38] was used in this study. Since in some of the previous studies, the most common kernels that obtained the best improvements were

Table 1 Descriptive statistics of heavy metals along with the associated standards and the calculated soil pollution index (SPI)

	Ag (mg/kg)	Co (mg/kg)	Pb (mg/kg)	Tl (mg/kg)	Be (mg/kg)	Ni (mg/kg)	Cd (mg/kg)	Ba (mg/kg)	Cu (mg/kg)	V (mg/kg)	Zn (mg/kg)	Cr (mg/kg)	Pn
Mean	0.17	10.81	19.20	0.86	1.08	43.09	0.31	307.56	27.14	81.86	78.77	91.49	0.95
Standard deviation	0.07	3.72	9.48	0.34	0.30	69.71	0.29	78.14	14.92	23.18	27.74	71.09	0.96
Min	0.10	2.53	2.89	0.18	0.24	11.49	0.10	80.00	6.98	22.55	14.07	16.48	0.31
Max	0.50	30.92	87.74	2.16	3.47	644.94	1.47	663.73	133.57	193.15	210.49	739.35	9.23
Standard	4	40	50	5	5	50	1	300	100	100	200	110	–

RBF, polynomial and linear, while other known kernels achieved poor results, so these kernel functions were tried in this study.

Besides kernel function, the performance of SVMs directly depends on the support vector machine's parameters such as regulation parameter (C), insensitive loss function (ε), gamma parameter (γ) etc., and the sensitivity of results is based on the precise optimization of these parameters. Having optimized the associated parameters for each kernel function, eighty percents of the original data were selected randomly as the training set and the model was trained using the optimized parameters. Finally, the prediction of SPI was implemented using both the training data (e.g., re-substitution error) and test data (e.g., generalization error).

As stated earlier, the sensitivity of the results of SVMs hinges on the value of each parameter, so these parameters have to be optimized. There are three parameters for Radial Basis Function (RBF) kernel [39]: C (penalty parameter) and γ (a tuning parameter controlling the width of the kernel function) and epsilon value (ε). For linear SVMs, the penalty parameter (C) was the only optimized parameter and for polynomial kernel, the degree of polynomial was tuned. A good way of choosing the value of d (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 [40]. The cross-validation procedure can prevent the over-fitting problem. Over-fitting occurs when a forecasting model has good performance on the training data but its generalization ability (e.g., its performance on the testing data set) is poor.

The parameters to optimize in each experiment were encoded in a vector, bound to maximum and minimum values and tuned with a program written in MATLAB (R2013b). To have an independent data set for which the out-of-sample generalization error of the method is considered, five-fold cross-validation was applied on the training data.

One of the other modeling procedures utilized for the prediction of SPI was artificial neural network. To keep within the scope of this paper, we limited our survey of ANN models to the feed-forward neural network with one hidden layer. As a whole, too many hidden nodes may lead to the problem of over-fitting, whereas too few nodes in the hidden layer may cause the problem of under-fitting [41]. The linear transfer function (e.g., $y_i = x_i$) and the following transfer function was used for the output and hidden layers, respectively:

$$y_j = \tanh\left(\sum_{i=1}^{d} w_{ij}x_i + b_j\right) \tag{3}$$

where w_{ij} and b_j are the weight and bias parameters in which "i" and "j" subscripts refer to the input and neuron,

respectively. In addition, Levenberg–Marquardt algorithm was used to update the weight and bias of the network according to this formula:

$$x_{k+1} = x_k - \left[J^{\mathrm{T}}J + \mu I\right]^{-1} J^{\mathrm{T}} e \tag{4}$$

where J is the Jacobian matrix containing first derivatives of the network errors with respect to the weights and biases, e is a vector of network errors, I is the identity matrix, x is a vector containing weights and biases, and μ is a scalar value, respectively. Prior to the data introduction to the neural network, standardization of the data (i.e., the data have zero mean and unit standard deviation) was done according to the following equation:

$$Z_i = (x_i - \bar{x}_i)/s_i \tag{5}$$

In which, $\bar{x}_i$ and s_i are the mean and standard deviation of the observed variables, respectively, whereas Z_i is the standardized value. In this study, different hidden node sizes ranging from 5 to 40 were applied, and given the optimum number of hidden nodes (based on the minimum MSE), the best performed ANN was used for out-of-sample SPI prediction.

To reduce the risk of over-fitting which is a common problem in ANN modeling especially when the number of observations is small in comparison with that of features, early stopping [42] algorithm were utilized. To be comparable with the results of SVMs, an independent data set containing 80 percents of the original data was trained 20 times with different data divisions to training, validation and testing set. At the next step, the generalization ability of the ANN was considered on the rest of the data set. Since different random initial weights may produce different training results, thus the training over subsamples was performed at a fixed seed value [43]. The mean squared error (MSE) for both the 80 percents of the original data (training data) and the independent data (test data) was worked out, and the average MSE was regarded as the out-of sample generalization error in early stopping method.

It should be noted that, by application of SVM and ANN for the prediction of SPI, we do not intend to undermine the direct calculation of this index since most of the offered formula for the calculation of soil pollution indices (e.g., enrichment factor, contamination factor, geoaccumulation index) are simply enough to apply directly; however, since these indices have some shortcoming for example in the case of Nemerow's index applied in this study, the influence of maximum value in calculations has been overemphasized and the weight of factors has not been taken into account as well. Therefore, sometimes some modifications are necessary to obviate these disadvantages such as introduction of entropy to calculate weights etc., making the problem more complex than usual situations and may incur unintentional errors during sub-index calculations. In these cases, the application of modeling procedures like SVM and ANN would be more beneficial. In this research, we used the basic Nemerow's index formula to simplify the problem.

On the other hand, data analysis was done using the geostatistical methods, as described by Isaaks and Srivastava [44] and Goovaerts [45]. In linear geostatistics method, a normal distribution for the variable is desired in order to avoid distortions of data and low level of significance. In this study, the distribution of the data was tested for normality by the Kolmogorov–Smirnov (K–S) test. The logarithm transformation was performed on SPI for further analysis since these raw data sets did not follow a normal distribution pattern. Semivariogram model selections and model cross-validation were also done using the methods of Goovaerts [45]. Semivariogram was used to quantify the spatial variability of a regionalized variable, which relates dissimilarity between paired data values to the distance between each sample pair [44, 45]. The GS+(v.5.1) software was used to perform the ordinary kriging method, and mapping was done using the ArcGIS 10.1 software.

3 Results and discussion

The descriptive statistics of the analyzed heavy metals and the calculated SPI along with the associated Iranian standards for heavy metals in soil published by the Iranian Department of Environment have been given in Table 1.

Considering this table, the mean value of Ba (307.56 mg/kg) is higher than that of the assigned standard value. This element varied from 80 to 663.73 mg/kg which was roughly in the same range as that of Eriksson [46] in the agricultural soils of Sweden (383–778 mg/kg) but higher than the values of Brazilian's soils (32.86–128.89 mg/kg) [47] and that was found in the soils of Buffalo, USA (50.9–553 mg/kg) [48]. The mean values of Pb, Zn, Ni, Co, Cd were lower than that reported by Esmaeili et al. [49]. in the industrial zone of Isfahan, Iran, which were 34.6, 111.5, 66.2, 14.7, 0.43 mg/kg, respectively. However, the average value of Cr (91.49 mg/kg) is higher than the mean detected value (85.9 mg/kg) in the latest study. Moreover, the mean values of Zn, Cu, Ni and Cr in this research were higher than that in the soils of China (in turn 58.9, 18.9, 20.8, 49.7 mg/kg) [50]. In addition, for other heavy metals like Cr, Ni and V, the mean concentrations are roughly near their standard values indicating a possible high risk associated with these heavy metals in short term. As it is obvious, the maximum values for most of these heavy metals have exceeded the standard level showing the gross pollution in some parts of the study area. For instance, level as high as 739.35 mg/kg has been detected for Cr which is roughly more than six times higher than that of the standard value. On the contrary, the calculated Nemerow's

synthetical pollution index (P_n) has been rendered in the last column of Table 1. With respect to this table, and the classification criterions for polluted index of soil (Table 2), it can be concluded that on average most of the study area is located in precaution domain, whereas the maximum value as high as 9.23 shows that some parts are seriously polluted with heavy metals. The study area has been classified given this index, and the result has been illustrated in Fig. 1. This figure shows that the right side of the study area is the most polluted part.

Referring to geological formations of the study area, the main lithologic units of this area are ophiolitic complex accompanied by Eocene–Oligocene volcaniclastic and basic rock units. The ophiolitic complex is the main body of ultramafic rocks. It has been proved that high concentrations of some elements such as Cr and Ni are due to presence of ultramafic rocks [51]. As mentioned earlier, the concentration of the above-mentioned heavy metals is near their standard values implying their possible geological source. On the other hand, mining activities and coal washing which are prevalent in the area (Fig. 2) are other sources that can be attributed for the elevated level of some heavy metals. In this field, Ardejani et al. [52]. in their study on Alborz Sharghi coal washing plant located about 55 km of Shahrood City, reported elevated levels of Fe, Mn, Zn, Cr and Co at a depth of 2 m from the top soil in the vicinity of this plant.

Table 2 Classification criterion for soil pollution index

Grade	Synthetical index	Appraisal result
1	$P_n \leq 0.7$	Safety domain
2	$0.7 < P_n \leq 1.0$	Precaution domain
3	$1.0 < P_n \leq 2.0$	Slightly polluted domain
4	$2.0 < P_n \leq 3.0$	Moderately polluted domain
5	$P_n > 3.0$	Seriously polluted domain

Adopted from [37]

Since the performance of learning machines is influenced by the parameter dimensions, so in the current research, the sensitivity of SVMs to the input parameters was considered. For the case of ANNs, the only considered parameter was the number of hidden nodes which has great impact on the predictive ability of the ANNs [43, 53, 54]. The results of the related parameters for linear, RBF and polynomial kernels have been presented in Tables 3, 4, 5, 6 and 7, respectively.

Considering Table 3, the best value of regulation parameter (C) for linear kernel was 3. Tuning of this parameter resulted in MSE of 0.014 and 0.017 for the training data, while R^2 of 0.985 and 0.988 for the training and test data was obtained during this process. On the contrary, the optimal values of ε, γ and C were 7, 0.2 and 0.00001, respectively (Tables 4, 5, 6, 7). Finally, the best

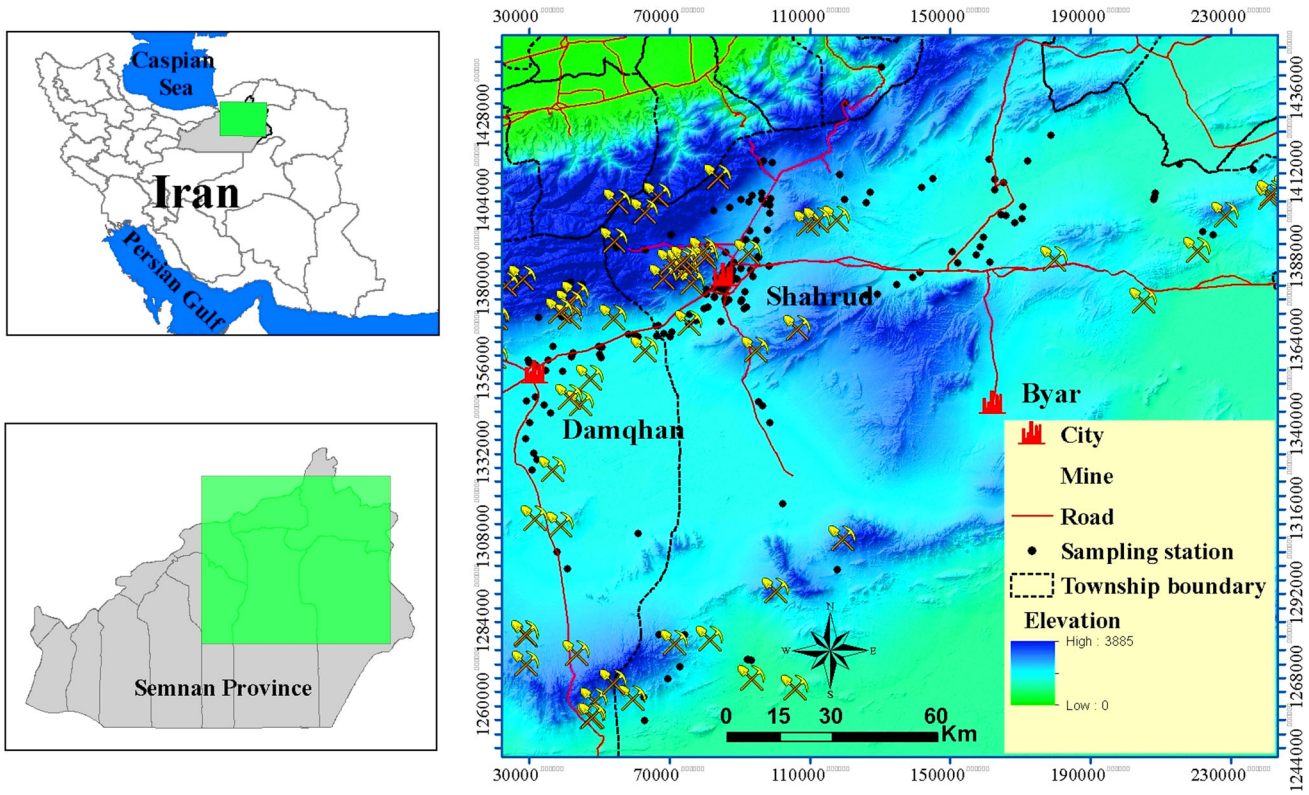

Fig. 1 The results of interpolation of the soil pollution index in the study area using geostatistical methods

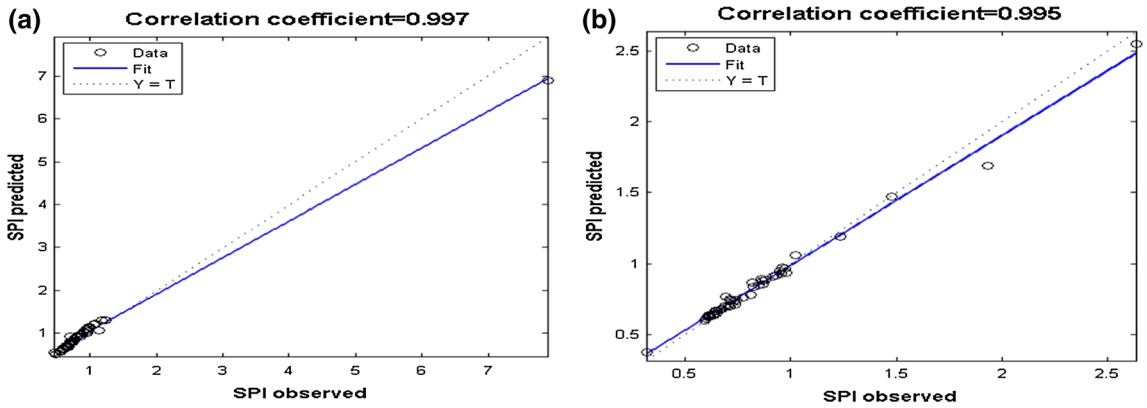

Fig. 2 The observed versus predicted values of soil pollution index (SPI) for modeling with RBF kernel (**a**) and ANNs with early stopping (**b**)

Table 3 The results of parameter optimization for linear kernel SVMs (–*p* stands for epsilon value)

$-p = 0.0625$ Regulation parameter (C)	Training MSE	Training R^2	Testing data MSE	Testing data R^2
1	0.011	0.991	0.008	0.955
2	0.013	0.989	0.012	0.922
3	0.014	0.985	0.017	0.988
4	0.054	0.954	0.030	0.777
5	0.019	0.983	0.034	0.753
6	0.058	0.963	0.017	0.642
7	0.262	0.781	0.513	0.763
8	0.228	0.930	0.299	0.939
9	1.460	0.890	0.363	0.749
10	0.899	0.451	1.162	0.636
11	0.212	0.843	0.128	0.126
12	0.107	0.919	0.243	0.207
13	0.418	0.782	0.278	0.190
14	0.570	0.724	0.657	0.143
15	0.672	0.608	0.859	0.719

Table 4 The results of optimization of epsilon value for RBF kernel SVMs (epsilon, –*g*, and regularization parameter, –*C*, were set to their default values)

$-g = 0.00015$, $-C = 5$ Epsilon value (P)	Training data MSE	Training data R^2	Testing data MSE	Testing data R^2
0	0.0318	0.985	0.552	0.839
0.1	0.030	0.984	0.504	0.882
0.2	0.017	0.993	0.026	0.823
0.3	0.041	0.991	0.113	0.643
0.4	0.098	0.992	0.127	0.825
0.5	0.159	0.988	0.247	0.738
0.6	0.265	0.989	0.309	0.727
0.7	0.375	0.988	0.409	0.811
0.8	0.527	0.983	1.043	0.849
0.9	0.693	0.983	0.704	0.862
1	0.857	0.990	0.915	0.703

Table 5 The results of optimization of gamma parameter for RBF kernel

$-p = 0.2, -C = 5$ Gamma parameter (g)	Training data MSE	Training data R^2	Testing data MSE	Testing data R^2
0	0.136	0.976	1.120	0.004
0.00001	0.012	0.991	0.050	0.967
0.0001	0.017	0.994	0.027	0.780
0.001	0.111	0.977	0.068	0.561
0.01	0.149	0.984	0.159	0.102
0.1	0.153	0.981	0.151	0.042
1	0.152	0.981	0.132	0.026

Table 6 The results of optimization of regularization parameter for RBF kernel

$-g = 0.00001, -p = 0.2$ Regulation parameter (C)	Training data MSE	Training data R^2	Testing data MSE	Testing data R^2
1	0.252	0.901	0.381	0.931
2	0.059	0.965	0.182	0.960
3	0.025	0.991	0.063	0.991
4	0.012	0.990	0.108	0.949
5	0.007	0.994	0.019	0.990
6	0.009	0.996	0.007	0.969
7	0.004	0.995	0.016	0.992
8	0.005	0.994	0.009	0.995
9	0.006	0.995	0.007	0.868
10	0.019	0.993	0.030	0.751

Table 7 The results of optimization of polynomial degree for polynomial kernel

$-g = 0.00001, -p = 0.2, -C = 5, -r = 0$ Polynomial degree (d)	Training data MSE	Training data R^2	Test data MSE	Test data R^2
1	0.017	0.988	0.020	0.847
2	0.017	0.992	0.015	0.800
3	0.009	0.991	0.106	0.984
4	0.009	0.993	0.009	0.916
5	0.009	0.994	0.120	0.745
6	0.012	0.989	0.883	0.348
7	0.011	0.991	0.043	0.930
8	0.013	0.990	0.059	0.837
9	0.015	0.989	2.454	0.475
10	0.014	0.987	8.144	0.489

$-r$ parameter of the kernel projection

polynomial degree was 4 for the polynomial kernel method. The results of parameter optimization show that the most sensitive parameter for the generalization ability of SVMs is γ. For instance, according to Table 5, by changing the value of γ from 0.00001 to 0.01, the R^2 of the test data set would reduce from 0.97 to 0.1. Since this parameter controls the amplitude of the kernel function, so the generalization ability of kernel hinges on it [55]. According to the previous studies, the regularization parameter (C) controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data [56, 57]. However, the results of this study showed that this parameter is not as important as that of gamma on the out-of-sample generalization of SVMs. On the other hand, ε-Insensitivity prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution [55]. Although the performance of the three kernel functions was not that much different; however, since the best results have been obtained for the polynomial kernel, using

Table 8 The results of ANNs with different number of hidden nodes

Number of hidden nodes	Average MSE for training set	Average MSE for testing set
5	0.09	0.13
10	0.14	0.07
15	0.07	0.01
20	0.10	0.02
25	0.07	0.17
30	0.05	0.15
35	0.04	0.14
40	0.01	1.91

the optimized parameters, the SVM was trained and tested with this kernel resulted in 15 support vectors (the points outside the ε-tube) out of 183 training samples.

On the contrary, the results of training with ANNs using early stopping (Table 8) indicate that the best generalization ability belongs to a neural network with 15 hidden nodes. As the number of hidden nodes increased, the generalization error decreased and for a neural network with 40 hidden nodes the average MSE of the testing set augmented to 1.91 compared with 0.01 for 15 hidden nodes, implying an obvious over-training of the model.

As a whole, the results of ANNs are comparable with that of SVMs, however, the generalization error of SVMs is quite a bit better than that of ANNs with early stopping. The same results have been obtained by other researchers through comparison the performance of SVMs with ANNs [29, 55, 58]. To graphically show the performance of these two methods, the predicted values for each method have been plotted against that of the target SPI for the testing set and the results have been shown in Fig. 2 for RBF kernel's and ANNs, respectively. The correlation coefficients for the RBF kernel and ANN with early stopping were 0.997 and 0.995, indicating roughly the same performance of these two methods. Our results are in accordance with that of Dibike et al. [59].

High dimensionality of the input space is often a serious problem associated with learning machines. A large training set that is able to provide a good distribution of high-dimensional data is essential for successful learning [43]. As the number of samples was significantly higher than the number of features (about 15 times that of features), so over-fitting due to the small data record was not a serious problem. One of the appealing features of SVMs is that the minimum found in the parameter space is always the global one [60]. That is to say, the problem of local minima which is common during training with ANNs is obviated in SVMs. Despite the different algorithms available for training a ANN, none of them can guarantee that the global rather a local minimum will be found in the training process [61]. Therefore, for cases in which there are quite comparable results for ANNs and SVMs, the usage of SVMs is preferable.

4 Conclusion

In this study, two learning machines (ANNs and SVMs) were evaluated and compared for predicting SPI with respect to the concentrations of heavy metals in a study area in Semnan Province, Iran. Since the number of samples was quite high (229 samples) in comparison with the number of features (12 heavy metals), so the two models could avoid the risk of over-fitting and their respective generalization ability was high and nearly the same, accordingly. As a whole, the results of ANNs were comparable with that of SVMs, however, the generalization error of SVMs is quite a bit better than that of ANNs with early stopping. Because of the fact that this is the first published literature on the usage of learning theory to model soil pollution, so, besides ANNs, SVMs can be an efficient modeling procedure in this field in feature studies.

Acknowledgments The authors are grateful to Geological Survey of Iran for the help in analysis of heavy metals. The financial support of this project has been provided by the Grant no. 100-2164 offered by Geological Survey of Iran.

Compliance with ethical standards

Hereby we confirm that there is no conflict of interest associated with this manuscript and all of the people who have contributed to the preparation of this paper have been cited by the authors. In addition, the funding agency related to this manuscript has also been cited in the acknowledgement section.

References

1. Gergen I, Harmanescu M (2012) Application of principal component analysis in the pollution assessment with heavy metals of vegetable food chain in the old mining areas. Chem Cent J 6:1–13
2. Loska K, Wiechula D, Barska B, Cebula E, Chojnecka A (2003) Assessment of arsenic enrichment of cultivated soils in Southern Poland. Pol J Environ Stud 2:187–192
3. Boszke L, Astel A (2009) Application of neural-based modeling in an assessment of pollution with mercury in the middle part of the Warta River. Environ Monit Assess 152(1–4):133–147
4. Sun Y, Zhou Q, Xie X, Liu R (2010) Spatial, sources and risk assessment of heavy metal contamination of urban soils in typical regions of Shenyang, China. J Hazard Mater 174:455–462
5. Ruf A (1998) A maturity index for predatory soil mites (Mesostigmata: Gamasina) as an indicator of environmental impacts of pollution on forest soils. Appl Soil Ecol 9:447–452
6. Atafar Z, Mesdaghinia A, Nouri J, Homaee M, Yunesian M, Ahmadimoghadam M, Mahvi AH (2010) Effect of fertilizer application on soil heavy metal concentration. Environ Monit Assess 160:83–89
7. Dankoub Z, Ayoubi S, Khademi H, Lu SH (2012) Spatial distribution of magnetic properties and selected heavy metals in

calcareous soils as affected by land use in the Isfahan region, Central Iran. Pedosphere 22:33–47
8. Jalali M, Khanlari ZV (2008) Effect of aging process on the fractionation of heavy metals in some calcareous soils of Iran. Geoderma 143:26–40
9. Saeedi M, Hosseinzadeh M, Jamshidi A, Pajooheshfar SP (2009) Assessment of heavy metals contamination and leaching characteristics in highway side soils, Iran. Environ Monit Assess 151:231–241
10. Cheng JL, Shi Z, Zhu YW (2007) Assessment and mapping of environmental quality in agricultural soils of Zhejiang Province, China. J Environ Sci 19:50–54
11. Chen SH, Jakeman AJ, Norton JP (2008) Artificial intelligence techniques: an introduction to their use for modelling environmental systems. Math Comput Simul 78(2–3):379–400
12. Hanrahan G (2011) Artificial neural network in biological and environmental analysis. Taylor and Francis Group, London, pp 119–147
13. Kisi O, Akbari N, Sanatipour M, Hashemi A, Teimourzadeh K, Shiri J (2013) Modeling of dissolved oxygen in river water using artificial intelligence techniques. J Environ Inform 22(2):92–101
14. May DB, Sivakumar M (2009) Prediction of urban storm water quality using artificial neural networks. Environ Model Softw 24(2):296–302
15. Nour MH, Smith DW, Gamal El-Din M, Prepas EE (2006) The application of artificial neural networks to flow and phosphorus dynamics in small streams on the Boreal Plain, with emphasis on the role of wetlands. Ecol Model 191(1):19–32
16. Ozkaya B, Demir A, Bilgili S (2007) Neural network prediction model for the methane fraction in biogas from field-scale landfill bioreactors. Environ Model Softw 22(6):815–822
17. Panda SS, Garg V, Chaubey I (2004) Artificial neural networks application in lake water quality estimation using satellite imagery. J Environ Inform 4(2):65–74
18. Wieland R, Mirschel W, Zbell B, Groth K, Pechenick A, Fukuda K (2012) A new library to combine artificial neural networks and support vector machines with statistics and a database engine for application in environmental modeling. Environ Model Softw 25:412–420
19. Aryafar A, Gholami R, Rooki R, Doulati Ardejani F (2012) Heavy metal pollution assessment using support vector machine in the Shur River, Sarcheshmeh copper mine, Iran. Environ Earth Sci 67:1191–1199
20. Gill MK, Asefa T, Kemblowski MW, McKee M (2006) Soil moisture prediction using support vector machines. J Am Water Resour Assoc 42:1033–1046
21. Guo Q, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. Ecol Model 182:75–90
22. Sadeghi R, Zarkami R, Sabetraftar K, 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
23. Haghverdi A, Cornelis WM, Ghahraman B (2012) A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. J Hydrol 442–443:46–54
24. Liao K, Xu S, Wu J, Zhu Q, An L (2014) Using support vector machines to predict cation exchange capacity of different soil horizons in Qingdao City, China. J Plant Nutr Soil Sci 177:775–782
25. Tamari S, Wösten JHM, Ruiz-Suárez JC (1996) Testing an artificial neural network for predicting soil hydraulic conductivity. Soil Sci Am J 60:1732–1741
26. Jiang H, Cotton WR (2004) Soil moisture estimation using an artificial neural network: a feasibility study. Can J Remote Sens 30(5):827–839
27. Yu Z, Liu D, Lu H, Fu X, Xiang L, Zhu Y (2012) A multi-layer soil moisture data assimilation using support vector machines and ensemble particle filter. J Hydrol 475:53–64
28. Were K, Bui DT, Dick OB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol Indic 52:394–403
29. Yoon H, Jun SH, Hyun Y, Bae GO, Lee KK (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
30. Dreyfus G (2005) Neural networks methodology and applications. Springer, Berlin, pp 1–493
31. Hsieh WM (2009) Machine learning methods in the environmental science, neural networks and kernels. Cambridge University Press, Cambridge, pp 86–157
32. Taylor BJ (2006) Methods and procedures for the verification and validation of artificial neural networks. Springer, Berlin, pp 1–275
33. Theodoros E, Tomaso P, Massimiliano P (2002) Regularization and statistical learning theory for data analysis. Comput Stat Data Anal 38:421–432
34. Peng C, Wen X (1999) Recent applications of artificial neural networks in forest resource management: an overview. Environmental decision support systems and artificial intelligence. Aaai Workshop, pp 15–22
35. Shokri BJ, Ramazi H, Doulati F, Moradzadeh A (2014) A statistical model to relate pyrite oxidation and oxygen transport within a coal waste pile: case study, Alborz Sharghi, northeast of Iran. Environ Earth Sci 71:4693–4702
36. IAEA (2004) Soil sampling for environmental contaminants, International atomic energy agency, Austria, 81 p
37. Liang CJ, Zhou S, Wei ZY (2007) Assessment and mapping of environmental quality in agricultural soils of Zhejiang Province, China. J Environ Sci 19:50–54
38. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm
39. Li H, Liang Y, Xu Q (2009) Support vector machines and its applications in chemistry. Chem Intell Lab Syst 95:188–198
40. Hoang H, Lock K, Mouton A, Goethals PLM (2010) Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. Ecol Inform 5:140–146
41. Khalil B, Ouarda TBMJ, St-Hilaire A (2011) Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. J Hydrol 405:277–287
42. Piotrowski AP, Napiorkowski JJ (2013) A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modeling. J Hydrol 476:97–111
43. Khalil A, Almasri MN, McKee M, Kaluarachchi JJ (2005) Applicability of statistical learning algorithms in groundwater quality modeling. Water Resour Res 41:1–16
44. Isaaks EH, Sivastava RM (1998) An introduction to applied geostatistics. Oxford University Press, New York
45. Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York
46. Eriksson JE (2001) Concentrations of 61 trace elements in sewage sludge, farmyard manure, mineral fertilizers, precipitation and in oil and crops. Swedish EPA Rep 5159, Stockholm
47. Juchen CR, Cervi EC, Boas MAV, Charlesworth S, Poleto C (2014) Comparative of local background values for trace elements in different Brazilian tropical soils. Int J Environ Eng Nat Resour 1(6):255–261
48. Cai M, McBride MB, Li K (2015) Bioaccessibility of Ba, Cu, Pb, and Zn in urban garden and orchard soils. Environ Pollut 208:145–152

49. Esmaeili A, Moore F, Keshavarzi B, Jaafarzadeh N, Kermani M (2014) A geochemical survey of heavy metals in agricultural and background soils of the Isfahan industrial zone, Iran. Catena 121:88–98
50. Sun C, Liu J, Wang Y, Sun L, Yu H (2013) Multivariate and geostatistical analyses of the spatial distribution and sources of heavy metals in agricultural soil in Dehui, Northeast China. Chemosphere 92(5):517–523
51. Kelepertsis A, Alexakis D, Kita I (2001) Environmental geochemistry of soils and waters of Susaki area, Korinthos, Greece. Environ Geochem Health 23:117–135
52. Ardejani FD, Shokri BJ, Moradzadeh A, Shafaei SZ, Kakaei R (2011) Geochemical characterisation of pyrite oxidation and environmental problems related to release and transport of metals from a coal washing low-grade waste dump, Shahrood, northeast Iran. Environ Monit Assess 183:41–55
53. Wanas NM, Auda G, Kamel M, Karray F (1998) On the optimal number of hidden nodes in a neural network. In: Proceedings of the IEEE Canadian conference on electrical and computer engineering, pp 918–921
54. Shu C, Ouarda TBMJ (2007) Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resour Res 43:1–12
55. Noori R, Karbassi A, Farokhnia A, Dehghani M (2009) Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. Environ Eng Sci 26(10):1503–1510
56. Ji AB, Pang JH, Qiu HJ (2010) Support vector machine for classification based on fuzzy training data. Expert Sys Appl 37:3495–3498
57. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, Fan BT (2004) Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 44:1257–1266
58. Zhang X, Srinivasan R, Van Liew M (2009) Approximating SWAT model using artificial neural network and support vector machine. J Am Water Resour Assoc 45(2):460–474
59. Dibike YB, Velickov S, Solomatine DP, Abott MB (2001) Model induction with support vector machines: introduction and applications. J Comput Civ Eng 15(3):208–216
60. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector methods. Cambridge University Press, Cambridge
61. Abraham A (2004) Meta learning evolutionary artificial neural networks. Neurocomputing 56:1–38