Journal of AI and Data Mining
Vol 6, No 1, 2018, 69-78

Ensemble of M5 Model Tree-Based Modelling of Sodium Adsorption Ratio

M. T. Sattari[1]*, M. Pal[2], R. Mirabbasi[3] and J. Abraham[4]

1. Department of Water Engineering, Agriculture Faculty, University of Tabriz, Tabriz, Iran.
2. Department of Civil Engineering, National Institute of Technology, Kurukshetra, 136119, Haryana, India.
3. Department of Water Engineering, Agriculture Faculty, University of Shahrekord, Shahrekord, Iran.
4. University of St. Thomas, School of Engineering, 2115 Summit Ave, St. Paul, MN 55105-1079, USA.

Received 21 March 2017; Revised 06 June 2017; Accepted 09 July 2017
**Corresponding author: mtsattar@gmail.com (MT. Sattari).*

Abstract
In this paper, we report the results of four ensemble approaches with the M5 model tree as the base regression model to anticipate Sodium Adsorption Ratio (SAR). The ensemble methods that combine the output of multiple regression models have been found to be more accurate than any of the individual models making up the ensemble. In this work, the additive boosting, bagging, rotation forest, and random sub-space methods were used. A dataset consisting of 488 samples with nine input parameters was obtained from the Barandoozchay River in the West Azerbaijan province, Iran. The three evaluation criteria correlation coefficient, root mean square error, and mean absolute error were used to judge the accuracy of different ensemble models. In addition to the use of an M5 model tree as the learning algorithm to predict the SAR values, a wrapper-based variable selection approach and a genetic algorithm were also used to select useful input variables. The encouraging performance motivates the use of this technique to predict the SAR values.

Keywords: *Water Quality, Sodium Adsorption Ratio, Data Mining, M5 model tree, Genetic Algorithm, Iran.*

1. Introduction

Surface water quality assessment and control is one of the important issues in water resource planning and management, especially in countries like Iran. It is a necessary step for the development of agricultural land, design and operation of irrigation systems, and crop pattern selection in the region. Since Iran contains arid and semi-arid regions and is faced with water deficiencies, an accurate estimation of parameters for water resource quality is necessary to make water resource decisions.

Rivers, being a major source of water, are important due to their role in providing the required amount of water for industrial, agricultural, and drinking uses. Release of urban, industrial, and agricultural wastes may change water quality within the rivers. Consequently, monitoring the water quality in various parts of rivers is very important. The water quality parameters such as pH, total dissolved solids (TDS), electrical conductivity (EC), and sodium adsorption ratio (SAR) are used to evaluate the quality of water for irrigation. Out of the different parameters used to assess irrigation water quality, SAR is important; its accurate estimation and assessment is necessary for water planning for agricultural uses.

In the recent years, several methods including statistical models, deterministic and machine learning models, especially artificial neural networks (ANN), M5 model trees, and support vector regressions have been proposed to analyze and model the water quality with the available data [1-4]. Most of these studies suggest that an improved performance is achieved by various machine-learning approaches in modelling water quality when compared to the statistical methods.

Performance of a machine-learning algorithm can further be improved by combining the outputs of several individual regression models. Each one of the individual models provides a solution to the same task [5, 6]. The approach of combining the

outputs of several regression models is known as the ensemble method. The ensemble approaches are capable of predicting a response by aggregating predictions from different trained regression models, which often results in the ensemble providing a better discernment than any individual model. Methods of creating ensemble of regression models proposed in the literature [7] include (1) manipulating the training data with a single algorithm, (2) use of different training parameters with a single algorithm, (3) use of different subsets of input variables with a single algorithm, and (4) combining outputs of different algorithms with the same training data.

Even though different approaches of creating ensemble models have been proposed in the literature, there are no clear guidelines about which method is the best. As was the case in the selection of the data, performance of a group of events might depend on the way the base algorithm was selected. Variable selection is the process of selecting a subset of input variables among the main data. This process reduces the number of variables to be used by a regression model allowing regression algorithms to operate faster, and provide a more compact model and possibility of improved function [8].

The design of an efficient and error-resistant variable-selection algorithm is an important step in water resource applications [9]. Based upon the criterion of whether a regression algorithm is used to evaluate the subsets of variables or not, the variable-selection methods can be divided into three categories, namely the filter, wrapper, and embedded methods [8-10]. The filter methods use an evaluation function in combination with a search method to select a subset of variables. The wrapper methods use a search algorithm and a regression model to search through the space of possible variables and select a subset of variables by directly optimizing its function (e.g. root mean square error) with the model used. In contrast to the filter and wrapper approaches, the embedded approach for variable selection is built into the regression/classification algorithm itself, and the prediction and variable selection process cannot be separated. Detailed discussions about variable selection in water resource applications have been presented in [11] and [12].

Keeping in view the improved performance of ensemble approaches for various applications [13-16], this paper proposes to use four ensemble approaches (two based on manipulating the training data with a single algorithm and two based on using different subsets of input variables with a single algorithm) with the M5 model tree. A tree-based regression approach is used for SAR values of the Barandoozchay River. Further, this work also proposes a wrapper-based variable selection approach to judge its effectiveness in predicting the SAR values. For the wrapper approach, a genetic algorithm (GA) and an M5 model tree are used as the search and regression algorithms, respectively.

Recently, researchers have tried to optimize different factors for the development of a more accurate model. For instance, Wu et al. (2010) [17] have used a data-driven model for rainfall forecasting and have optimized the model from the inputs, methods, and data-preprocessing viewpoints. Rain data records from four different stations that included both monthly and daily values were evaluated for this work. Also four different model, namely modular artificial neural network (MANN), artificial neural network (ANN), K-nearest-neighbors (K-NN), and linear regression (LR), were compared with each other in order to find the best performance. The results obtained for the models show that MANN performs the best among other models and can be quite suitable, especially for daily rainfall forecasting.

Chau and Wu (2010) [18] have used hybrid models for predicting daily rainfall. They included artificial neural networks (ANNs) and support vector regression (SVR). Two daily rainfall series were used to examine the accuracy of the model in forecasting 1-day-, 2-day-, and 3-day-ahead rainfall. The results obtained showed that the hybrid SVR model performed the best and decreased the errors.

Wang et al. (2015) [19] have presented a new method using the auto-regressive integrated moving average (ARIMA) model coupled with the ensemble empirical mode decomposition (EEMD) for the prediction of annual run-off. Several regions were studied for developing the mentioned method, and the results obtained indicated that EEMD could successfully improve the accuracy ratio and that the proposed EEMD-ARIMA simulation was confirmed as an effective model for forecasting annual run-off time series.

Gholami et al. (2015) [20] have used ANN combined with dendrochronology (using tree-rings) to simulate groundwater level changes during the period from 1912 to 2013 in an alluvial aquifer located at the Caspian southern coasts of Iran. For this purpose, they utilized the tree-ring diameter and precipitation as the input parameters and the groundwater levels as the output. The

results of this work revealed that this model had a high degree of precision and could be an effective method in simulating the groundwater levels. In addition, they suggested that this method could be useful for the prediction and evaluation of drought.

Taormina and Chau (2015) [21] have used data-driven techniques for modeling streamflow. They reported that the appropriate input variables could help to improve the model accuracy. Accordingly, they used particle swarm optimization (PSO) and Extreme Learning Machines (ELM) to select the best input parameters, and then developed a fast and accurate model. The results obtained indicate that the proposed techniques were suitable for the rainfall–runoff modeling applications, and could increase the prediction accuracy.

Chen et al. (2015) [22] have employed a hybrid neural network (HNN) model in combination with three optimization algorithms, namely differential evolution (DE), artificial bee colony (ABC), and ant colony optimization (ACO), to find the optimal downstream river flow forecasting. The stated algorithms were used to determine the premise parameters of HNN. Furthermore, PSO was used to compare the performance of the forecasting models. The results obtained demonstrated that DE had the best performance among the other algorithms in forecasting the downstream river flow, and also had a reliable accuracy such as PSO.

Fadaei-Kermani et al. (2017) [23] have tried to predict drought occurrence, based on the standard precipitation index (SPI), using k-nearest neighbor modeling. The model was tested by using precipitation data of Kerman, Iran. Results showed that the model gives reasonable predictions of drought situation in the region.

Measurements of the water quality parameters periodically are time-consuming and very expensive. According to the previous research works in the world, other intelligent system estimation methods have presented good results but the other data mining methods such as tree models have not been examined. In this work, we applied the M5 evaluation and ensemble method performance in SAR estimation for the conditions in Iran.

2. Materials and methods

2.1. M5 model tree

The M5 tree model is a common decision tree with a linear regression function at the terminal nodes [24]. This tree implements the divide-and-conquer method, and, unlike the other decision-making trees that are used for predicting discrete classes, it is used for predicting continuous numerical variables. Model tree generation involves two stages. The first step includes setting a division criterion for establishing the decision-making tree, whereas the second step involved a pruning method for pruning the overgrown trees. The M5 tree model algorithm uses the standard deviation of the class as a whole to identify the node that expresses the desired standard error. It further calculates the expected reduction in the error as a result of testing each variable at that node. The standard deviation reduction (SDR) formula used in the design of M5 model tree can be represented by:

$$SDR = sd(K) - \sum \frac{|K_i|}{|K|} sd(K_i) \quad (1)$$

where, K represents the number of examples reaching the node, _ represents the number of samples containing the i^{th} output of the potential set, and SD is the standard deviation. SD of a child node is less than that of its parent node, thus increasing the purity of the child node [24]. Upon evaluating all the possible division cases, the M5 tree model design method selects the division point that maximizes the mean error reduction. This data division during the M5 model creation produces a large tree which may be the cause of over-fitting with testing data. In order to remove the problem of over-fitting, Quinlan [24] has suggested the use of a pruning method to prune back the over-grown tree. In general, this pruning method is obtained by substituting the sub-tree with linear regression functions. Quinlan [24] and Witten and Frank [25] have presented further details of the M5 tree model.

2.2. Genetic algorithm

A Genetic Algorithm (GA) is a heuristic search method, and is based upon the principles of evolution via natural selection. GA simulates the "survival of the fittest" principle across consecutive generations using a population of singular solutions. Singular solutions and variables act like chromosomes and genes, respectively. GA starts through a process of random generation of an initial population of chromosomes and computing the fitness of each chromosome in the population. Chromosome eligibility is used as a criterion for determining the selection probability for each singular chromosome in the current population. In the second stage, a new population is generated from the initial population using genetic operators such as cross-over and mutation, and repeating this

process until the final solution is achieved. GA-based variable selection works by using a population of individual chromosomes (i.e. solution) consisting of a subset of input variables. A root mean square error (RMSE) value is used as a fitness function to evaluate the individual chromosomes for their reproductive success during the variable selection process. Further details about GAs have been provided by Mitchell [26], and their use for feature selection has been discussed by Pal [27].

2.3. Ensemble approaches

The four ensemble approaches stochastic gradient boosting, bagging, rotation forest, and random subspace were used with the M5 model tree as a base regression model. A brief discussion about different ensemble approaches is provided in following sections. Further details of various ensemble approaches have been provided by Seni and Elder [7].

2.4. Stochastic gradient boosting

Friedman [28] has proposed a gradient boosting-based ensemble technique for the regression models. Stochastic gradient boosting works in a similar way as the other boosting methods [29]; it generalizes them by optimizing an arbitrary differentiable loss function. The proposed algorithm uses a base model to obtain the eligibility of those training set sub-samples that are randomly selected in each iteration. The size of the sub-sample used in each iteration is a user-defined parameter, and is taken as a fraction of the size of the total training dataset. Generally, a smaller fraction of training dataset introduces randomness into the model and helps prevent over-fitting. The use of a smaller fraction of training dataset makes speed the algorithm because the base regression model has to fit smaller datasets at each iteration.

For a training dataset $\{(x_i, y_i), I = 1, 2, \dots n\}$, where x_i is an input vector described by p features and y_i is an output variable used during training with n number of training samples and a base algorithm $\varphi(y,F(x))$. The model is initialized with a constant value,

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} \varphi(y_i, \gamma) \quad (2)$$

The so-called pseudo-residuals are calculated from:

$$r_{im} = -\left[\frac{\partial \varphi(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x_i)=F_{m-1}(x)} \quad (3)$$

where, $-r_i$ is the path of steepest decent, φ is the loss function, and m = 1, 2,……M, where M is the number of iterations an algorithm is run.

The parameter γ_m is then calculated by:

$$\arg\min_{\gamma} \sum_{i=1}^{n} \varphi(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \quad (4)$$

The model is then updated by:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (5)$$

where, $h_m(x)$ is the base learner (i.e. M5 model tree in this work).

2.5. Random sub-space

The quasi-random method proposed by Hu [29] comprises a group of methods used for developing a model from the modified training data. In order to generate this model, the random subsets of input variables were used in each iteration to modify the training data. In the quasi-random method, r characteristics are randomly selected from the p-dimensional training dataset $\{(x_i, y_i), i = 1, 2, \dots n\}$. In this ensemble approach, only the input variables are resampled, whereas no sampling is used with the training dataset. Therefore, the modified training data includes a random r-dimensional subset of the principal p-dimensional characteristic space. The algorithm is applied to this set of new data in the R-dimensional random subspaces, and many models are trained across these subsets that are randomly selected among all the variables (e.g. the random space). Ultimately, the outputs from these models are combined based on the majority vote. The number of random sub-spaces r and the iterations to be performed is decided by the user. Further details about this algorithm have been provided by Ho [30].

2.6. Bagging

Bagging [31], also known as "bootstrap aggregating", is another group of technique used to increase the accuracy and validity of a base regression model. It operates based on the training data manipulation rule applied in each iteration. This method collects the results obtained from *h* regression models that are developed based on *h* self-commissioning training samples. In this approach, each training set of the regression model is developed via self-commissioning sampling, which includes random selection and substitution of N samples, where N is the size of the main training set. Boot-strapped sampling involves repeating many of the original samples in the resulting training set, while the others may be

left out. In the case of bagging, the training set consists of about 67% of the data from the original training set, and leaves out about one-third of the original training samples in each iteration. These left-out samples are known as the out-of-bag samples. The learning algorithm generates a regression model from the sampled training data, and is allowed to run a number of times (i.e. h, a user-defined parameter) using a different boot-strapped sample each time. The final result is obtained by averaging the outputs of individual models built over each boot-strapped sample. Breiman [30] has discussed further details about bagging.

2.7. Rotation forest

Rodriguez *et al*. [32] have proposed a group of methods based on the rotation forest method, and used the characteristics obtained from rotation of the sub-spaces of the main dataset to improve the accuracy of the base regression/classification model. This algorithm acts similar to the quasi-random method by randomly dividing the input characteristic set into *P* characteristic subsets (one parameter is defined by the user), and subsequently, applying the main component analysis (e.g. characteristic extraction) to this characteristics subset. Therefore, *P* axis rotations must take place for creating new characteristics for the base regression method. Each regression model in the group is developed via a set of transformed data. All the main components remain in the data in order to maintain the diversity of the information that exists in the data. The key to a successful implementation of the rotation forest method lies in using the rotation matrix formed by the linear transform subsets.

2.8. Dataset and methodology

With the background of the various analysis methods now articulated, attention can be refocused on the specific problem under consideration. The Barandoozchay River originates from the western side of the Bonad Yanjool Mountains lying along the border region of Iran and Turkey. After entering the Urmia plain, this river is divided into many branches, and finally flows to Urmia Lake. A total of 488 measurements for nine hydro-chemical parameters at three hydrometric stations (Babarud, Bikaran, and Dizaj) were made during a period from 1992 to 2010. The hydro-chemical parameters include total dissolved solids (TDS), electrical conductivity (EC), pH value, HCO3, chlorine (Cl), SO_4, calcium (Ca), magnesium (Mg), and sodium (Na). The Sodium Adsorption Ratio (SAR) was calculated for all water samples based on the following formula, provided by the U.S. Salinity Laboratory (1954):

$$SAR = \frac{\left(Na^{+}\right)}{\sqrt{1/2\left[\left(Ca^{2+}\right)+\left(Mg^{2+}\right)\right]}} \quad (6)$$

Ion concentrations are expressed in milliequivalents per liter (mEq/L). Based on Eq. 6, changes in the concentration of calcium and magnesium due to sedimentation or dissolution of alkaline carbonates are the important factors controlling the SAR value [33]. The SAR values were used for predictive modelling using the M5 model tree algorithm. The geographical information about the location of hydrometric stations is provided in table 1, and the location map of the Barandoozchay River and the considered hydrometric stations are shown in figure 1.

Before using the regression models, the data was checked for homogeneity and presence of outliers using the run test and box and wicker plot, respectively, calculated using the Minitab software. The statistical characteristics and correlation coefficient between SAR and different hydro-chemical parameters of the river water samples are provided in tables 2 and 3, respectively. For the M5 model tree-based .regression approach, only one parameter, i.e. the minimum number of training examples to be allowed at a leaf node, needs to be determined for a given dataset. After several trials, a value of four numbers of training examples was found to perform well with the dataset used. The use of ensemble approaches also requires setting of different user-defined parameters. In order to have the uniformity in number of times, a base model was used (i.e. M5 model tree), an amount of 10 was chosen for all the four ensemble methods. The optimal values for the other parameters with different ensemble approaches are provided in table 4.

Table 1. **Detailed information of hydrometric stations.**

St. No.	St. Name	Longitude	Latitude	Height (m)
1	Bibakran	00-44-54	00-37-17	1570
2	Dizaj	00-45-04	00-37-23	1320
3	Babaruod	00-45-14	00-37-24	1285

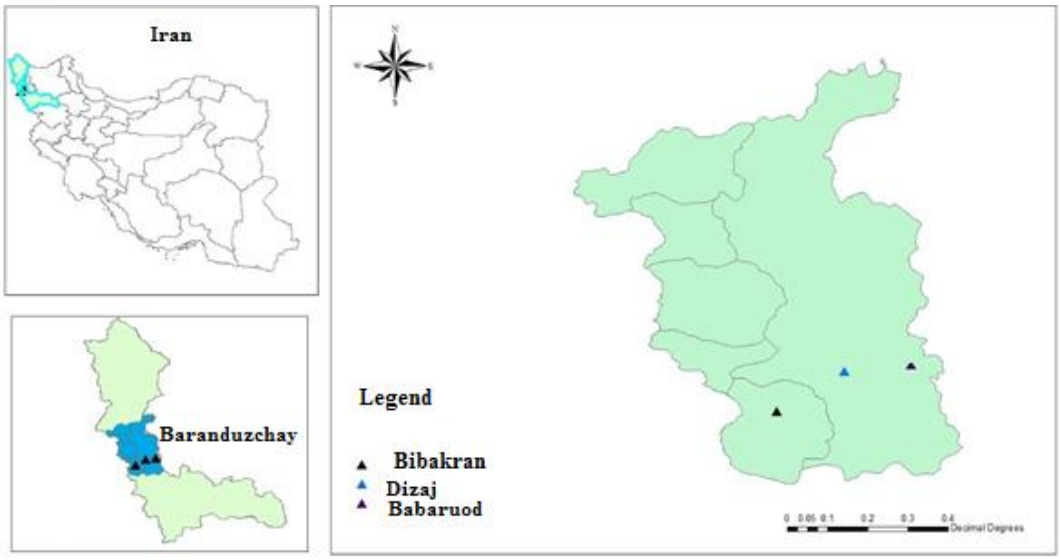

Figure 1. **Location map of Barandoozchay watershed and hydrometric stations.**

Table 2. Statistical characteristics of observed water quality data in Barandoozchay station.

Parameter	Minimum	Maximum	Mean	Standard Deviation	Unit
TDS	96.20	500.5	282.483	74.942	Mg/L
EC	148	770	434.589	115.296	mmhos/cm
pH	6.50	9.97	7.611	0.389	-
Cl	0.00	1.00	0.322	0.143	Mg/l
SO4	0.05	5.00	0.801	0.427	Mg/l
Ca	0.80	5.00	2.954	0.767	Mg/l
Mg	0.20	4.40	1.742	0.707	Mg/l
Na	0.10	1.20	0.300	0.145	Mg/l
HCo3	1.40	7.30	4.021	1.110	Mg/l
SAR	0.055	0.818	0.197	0.094	Mg/l

Table 3. Cross-correlation coefficient between SAR and other water quality parameters.

Parameter	Cross-Correlation
Na	0.954
SO_4	0.269
Cl	0.214
EC	0.207
Mg	0.121
pH	0.059
HCO3	0.055
Ca	-0.037
TDS	0.21

Table 4. User-defined parameters for different ensemble approaches.

Ensemble approach	User-defined parameters
Stochastic gradient Boosting	Shrinkage rate = 1
Random sub-space	Size of each subspace = 0.5 (if less than 1, percentage of the number of attributes)
Bagging	Size of each bag = 100% (i.e. percentage of training data in each iteration)
Rotation forest	Percentage of instances to be removed in each iteration = 50% P = 3

The use of GA as a search algorithm with the M5 model tree-based variable-selection approach requires setting parameters. After a large number of trials, the population size and number of generations = 50, probability of cross-over = 0.8, and probability of mutation = 0.033 were found to optimize the performance.

To judge the performance of the M5 model tree, a ten-fold cross-validation was used for model training and validation. Cross-validation is a method of estimating the accuracy of a regression model, where the input dataset is divided into several parts (say ten in ten-fold cross-validation), with each part, in turn, used to test a model fitted to the remaining parts. Thus the classification algorithm trains ten times with different datasets. Ultimately, the average error is obtained from the ten calculated error values. Three different statistical criteria including the correlation coefficient (CC), root mean square error (RMSE), and mean absolute error (MAE) were calculated for evaluating the success of different models in predicting the SAR values. The higher values of CC and the lower values of RMSE and MAE indicate a better performance by the model.

3. Results

Table 5 provides the correlation coefficient, RMSE, and MAE values using different ensemble approaches with the M5 model tree as a base regression model. The predictions by different ensemble approaches were compared with the actual SAR values and plotted in figure 2. The results tabulated in table 5 suggest an encouraging performance by the stochastic gradient boosting, bagging, and rotation forest-based ensemble approaches compared to the M5 model tree-based approach for the prediction of SAR values. The correlation coefficient value of 0.9978 (RMSE = 0.0062, MAE = 0.0031) indicates improved performance by stochastic gradient boosting-based approach in comparison with the other three ensemble approaches. These results also indicate a poor performance by random subspace-based ensemble even in comparison with the simple M5 model tree approach. A possible reason for this may be attributed to the choice of input variables selected during various runs of the random subspace-based ensemble approach.

Table 5. Values of different statistical measures obtained using different ensemble approaches with M5 model tree with different input variables.

Model used	With nine input parameters (TDS, Ec, pH, HCO3, Cl, SO_4, Ca, Mg, Na)			With 4 input variables (Cl, Ca, Mg, Na) selected by wrapper-based variable selection approach		
	CC	RMSE	MAE	CC	RMSE	MAE
M5 model tree	0.9960	0.0085	0.0041	0.9975	0.0067	0.0031
Stochastic gradient Boosting	0.9978	0.0062	0.0031	0.9986	0.0050	0.0023
Bagging	0.9966	0.0078	0.0038	0.9975	0.0067	0.0031
Random space	0.9446	0.0435	0.0282	0.9399	0.0489	0.0319
Rotation forest	0.9976	0.0066	0.0030	0.9980	0.0060	0.0027

The scatter plot displayed in figure 2 also justifies the improved performance by different ensemble approaches, except the random sub-space method, which under-predicts all the SAR values lying in the range of 0.25-1 and over-predicts all values lying between 0 and 0.25.

The results of the wrapper-based variable selection approach indicate that only four input variables (i.e. Cl, Ca, Mg, Na) are enough to predict the SAR values using the M5 model tree-based regression approach.

Thus to evaluate the performance of this combination of input variables in predicting the SAR values, different models using the M5 approach as well as all the four ensemble approaches were generated. Table 5 also provides the CC, RMSE, and MAE values achieved by different models with four selected input variables.

The results with the M5 model tree and various ensemble approaches using four selected variables indicate an improved performance by the reduced dataset in comparison wih the dataset consisting of all the nine variables, except in the case of the random subspace-based approach. The M5 model tree, when used with this reduced dataset, achieves better results in terms of CC, RMSE, and MAE (Table 5) in comparison with the datasets consisting of all the nine variables. The results from table 5 suggest that a stochastic gradient boosting-based ensemble approach again performs well in comparison with the other approaches with reduced dataset as well. The scatter plot between the actual and predicted SAR values using four input variables (Figure 3) also indicates that most of the predicted values by different approaches lie on the line of perfect agreement, thus justifying the larger values of correlation coefficient.

Further, the results of the random subspace-based ensemble approach with reduced dataset are similar to those obtained with the nine input variables, justifying that this approach cannot be suggested to predict the SAR values for the type of dataset used in this work.

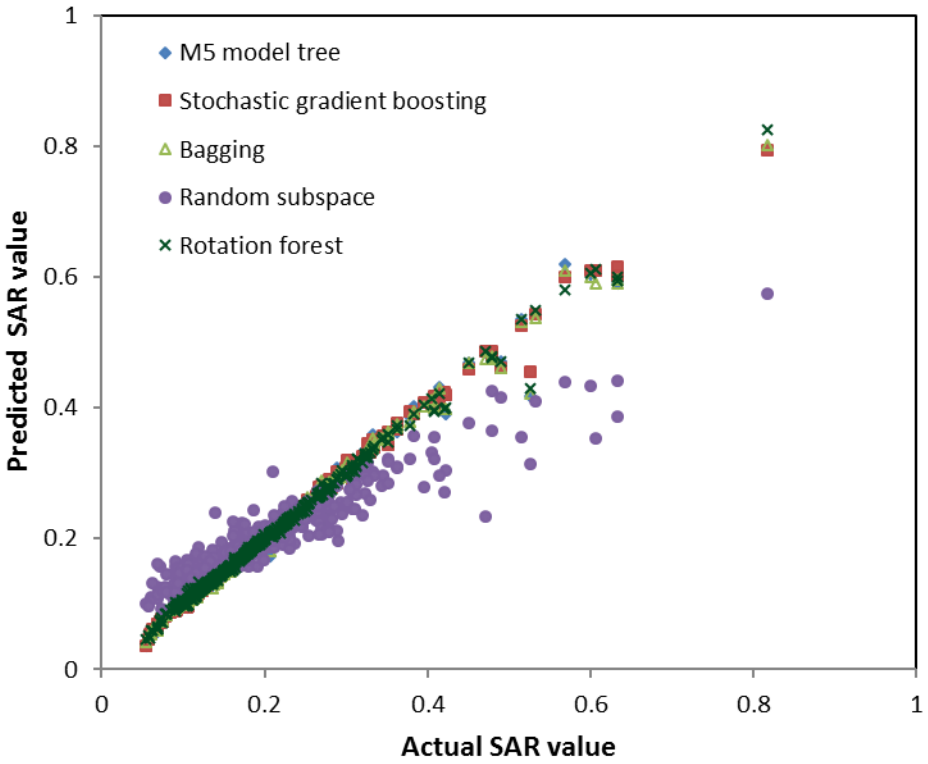

Figure 2. Actual vs. predicated SAR value by different ensemble approaches using M5 model tree and nine input variables.

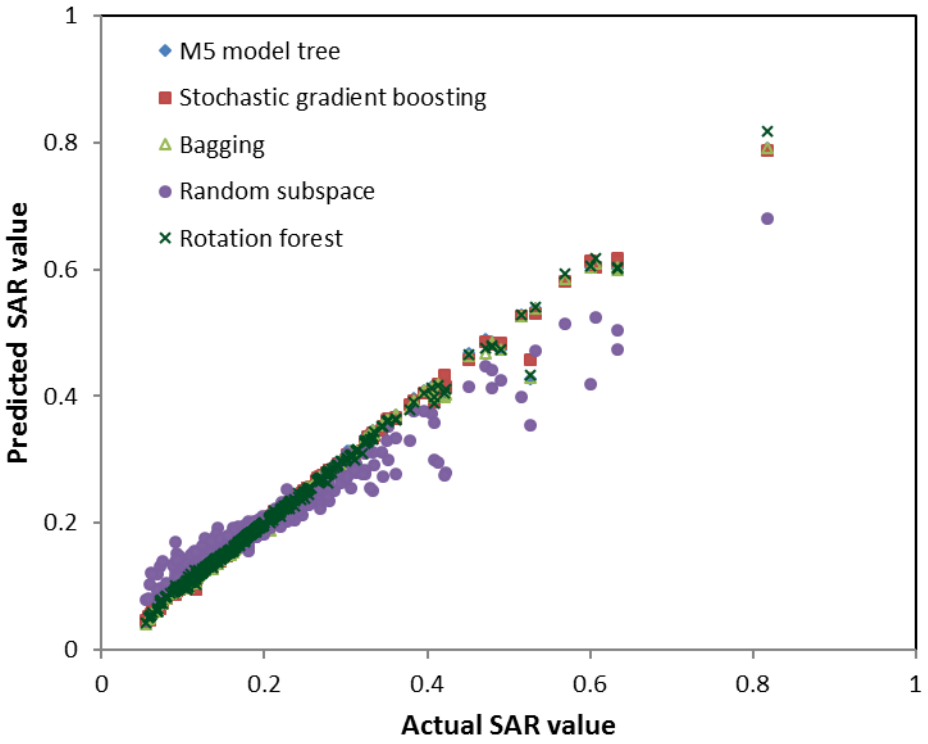

Figure 3. Actual vs. predicated SAR value by different ensemble approaches using M5 model tree and four input variables.

```
Na <= 0.355 :
|  Na <= 0.205 :
|  |  Na <= 0.145 :
|  |  |  Mg <= 1.65 : LM1
|  |  |  Mg >  1.65 : LM2
|  |  Na >  0.145 :
|  |  |  Ca <= 2.75 : LM3
|  |  |  Ca >  2.75 : LM4
|  Na >  0.205 :
|  |  Ca <= 2.65 : LM5
|  |  Ca >  2.65 : LM6
Na >  0.355 :
|  Na <= 0.475 : LM7
|  Na >  0.475 :
|  |  Na <= 0.655
|  |  |  Ca <= 3.5 : LM8
|  |  |  Ca >  3.5 : LM9
|  |  Na >  0.655 : LM10
```

LM1
SAR = 0.0002*Cl - 0.0135*Ca - 0.0138*Mg + 0.6935*Na + 0.0575
LM2
SAR = 0.0002*Cl - 0.0114*Ca - 0.0084*Mg + 0.6504*Na + 0.0487
LM3
SAR = 0.0002 * Cl - 0.0194 * Ca - 0.0174 * Mg + 0.7155 * Na + 0.0737
LM4
SAR = 0.0002 * Cl - 0.0129 * Ca - 0.0134 * Mg + 0.6517 * Na + 0.0625
LM 5
SAR = 0.004 * Cl - 0.0265 * Ca - 0.0245 * Mg + 0.7035 * Na + 0.1035
LM 6
SAR = 0.0066 * Cl - 0.0177 * Ca - 0.0179 * Mg + 0.6127 * Na + 0.0936
LM 7
SAR = 0.0032 * Cl - 0.026 * Ca - 0.0279 * Mg + 0.6133 * Na + 0.1455
LM 8
SAR = 0.0112 * Cl - 0.043 * Ca - 0.0389 * Mg + 0.6541 * Na + 0.1995
LM 9
SAR = 0.0112 * Cl - 0.0315 * Ca - 0.0324 * Mg + 0.6055 * Na + 0.1768
LM 10
SAR = 0.0193 * Cl - 0.0458 * Ca - 0.0471 * Mg + 0.6582 * Na + 0.214

Figure 4. Linear models for SAR prediction using M5 model tree with four input variables selected using variable selection approach.

Except performing well to predict the SAR values, the M5 model tree provides a simple linear relation between SAR and different input variables. Figure 4 provides the equations obtained by the M5 model tree using four input variables (Cl, Ca, Mg, Na) to predict the SAR values. These equations can easily be used to predict the SAR values by field engineers for the datasets lying within the ranges of the data used in this work.

4. Discussion and conclusions

The results obtained for this work demonstrate an encouraging performance by the proposed wrapper-based variable selection approach and different ensemble approaches to predict the SAR values. Except for the random sub-space approach, other ensemble approaches perform well in comparison with the M5 model tree algorithm. Another conclusion from this work is that the wrapper-based variable selection approach is able to reduce the number of variables required to predict the SAR values quite effectively. The reduced dataset was found to perform well in comparison with the dataset using all the nine input variables. In spite of improved performance by various ensemble approaches with the M5 model tree as the base algorithm, two issues, the increased computation cost and difficulty to interpret models, are still a challenge to researchers while dealing with ensemble learning. Furthermore, the choice of the base algorithm and the user-defined parameters required by different ensemble approach may affect the generalization capability of a base algorithm, suggesting the need for more research works with other base algorithms such as neural network and support vector machines.

The quality of the data mining technique is seen to depend on reliable and sufficient input data. With appropriate inputs, it is possible to monitor the quality of parameters such as SAR for a suitable management, and thereby, avoid negative impacts such as by pollution. A result is improvement in the social use of water resources in areas such as the Urmia lake basin.

References

[1] Nourani, V., RezapourKhanghah, T. & Sayyadi, M. (2013). Application of the Artificial Neural Network to monitor the quality of treated water. International Journal of Management & Information Technology, vol. 3, no. 1, pp. 38-45.

[2] Palani, S., Liong, S. Y. & Tkalich, P. (2008). An ANN application for water quality forecasting. Mar. Pollut. Bull, vol. 56, no. 9, pp. 1586-1597.

[3] Saghebian, S. M., Sattari, M. T., Mirabbasi, R. & Pal, M. (2014). Ground water quality classification by decision tree method in Ardebil region, Iran. Arabian J. Geosci, pp. 4767- 4777.

[4] Singh, K. P., Basant, A., Malik, A. & Jain, G. (2009). Artificial neural network modeling of the river water quality-A case study. Ecol. Modell, vol. 220, no. 6, pp. 888–895.

[5] Anctil, F. & Lauzon, N. (2004). Generalization for neural networks through data sampling and training procedures with applications to streamflow predictions. Hydrol. Earth Syst. Sci. Discuss, vol. 8, no. 5, pp. 940-958.

[6] Slomatine, D. P. & Seik, M. B. (2006). Modular learning models in forecasting natural phenomena. Journal Neural Networks special issue: Earth sciences and environmental, vol. 19, no. 2, pp. 215-224.

[7] Seni, G. & Elder, J. (2010). Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions. San Rafael, CA: Morgan and Claypool.

[8] Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, vol. 3, pp. 1157-1182.

[9] Galelli, S. & Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modeling. Water Resour. Res., vol. 49, no. 7, pp. 4295-4310.

[10] Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S.B. & Pintelas, P.E. (2007). Feature selection for regression problems. Proceedings of HERCMA'07.The eighth Hellenic European research on computer mathematics and its applications conference, Athens, Greece, 2007.

[11] May, R., Dandy, G. & Maier, H. (2011). Review of input variable selection methods for artificial neural networks, Available: http://cdn.intechopen.com/pdfs/14882/InTech_Review of_inputvariableselectionmethodsforartificialneuralnetworks.pdf (accessed on 19-07-2015).

[12] Ssegane, H., Tollner, E. W., Mohamoud, Y. M., Rasmussen, T.C. & Dowd, J.F. (2012). Advances in variable selection methods I: Causal selection methods versus stepwise regression and principal component analysis on data of known and unknown functional relationships. J. Hydrol., vol. 438, pp. 16-25.

[13] Erdal, H. I & Karakurt, O. (2013). Advancing monthly stream flow prediction accuracy of CART models using ensemble learning paradigms. J. Hydrol., vol. 477, pp. 119–128.

[14] Erdal, H. I., Karakurt, O. & Namli, E. (2013). High performance concrete compressive strength forecasting using ensemble models based on discrete wavelet transform. Eng. Appl. Artif. Intell., vol. 26, no. 4, pp. 1246-1254.

[15] Erdal, H. I. (2013). Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. Eng. Appl. Artif. Intell., vol. 26, no. 7, pp. 1689-1697.

[16] Rajagopalan, B., Grantz, K., Regonda, S., Clark, M. & Zagona, E. (2005). Ensemble streamflow forecasting: Methods and applications. Advances in Water Science Methodologies, pp. 97-116.

[17] Wu, C. L., et al. (2010). Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques, Journal of Hydrology, vol. 389, no. 1-2, pp. 146-167.

[18] Chau, K. W. & Wu, C. L. (2010). A Hybrid Model Coupled with Singular Spectrum Analysis for Daily Rainfall Prediction, Journal of Hydroinformatics, vol. 12, no. 4, pp. 458-473.

[19] Wang W. C., et al. (2015). Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition, Water Resources Management, vol. 29, no. 8, pp. 2655-2675.

[20] Gholami V., et al. (2015). Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers, Journal of Hydrology, vol. 529, no. 3, pp. 1060-1069.

[21] Taormina R., et al. (2015). Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines, Journal of Hydrology, vol. 529, no. 3, pp. 1617-1632.

[22] Chen X. Y., et al. (2015). A comparative study of population-based optimization algorithms for downstream river flow forecasting by a hybrid neural network model, Engineering Applications of Artificial Intelligence, vol. 46, no. PA, pp. 258-268.

[23] Fadaei-Kermani E., et al. (2017). Drought Monitoring and Prediction using K-Nearest Neighbor Algorithm, Journal of AI and data mining, vol. 5, no. 2, pp. 319-325

[24] Quinlan, J. R. (1992). Learning with continuous classes, Proceedings of Australian Joint Conference on Artificial Intelligence, Singapore: World Scientific Press, pp. 343–348.

[25] Witten, IH. & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.

[26] Mitchell, M. (1998). An introduction to genetic algorithms. MIT press.

[27] Pal, M. (2013). Hybrid genetic algorithm for feature selection with hyperspectral data. IEEE Geosci. Remote Sens., vol. 4, no. 7, pp. 619-628.

[28] Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, vol. 38, pp. 367-378.

[29] Freund, Y. & Schapire, R. E. (1996). Experiments with new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference (edited by Lorenza Saitta) CA: Morgan Kaufmann, San Fransisco: pp. 148-156.

[30] Rajagopalan, B., Grantz, K., Regonda, S., Clark, M. & Zagona, E. (2005). Ensemble streamflow forecasting: Methods and applications. Advances in Water Science Methodologies, pp. 97-116.

[31] Breiman, L. (1996). Bagging Predictors. Machine Learning, vol. 24, no. 2, pp. 123-140.

[32] Rodriguez, J. J., Kuncheva, L. I. & Carlos, J. (2006). Rotation forest: A new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell, vol. 28, no. 10, pp. 1619-1630.

[33] Albiac, J. & Dinar, A. (2012). The management of water quality and irrigation technologies. Earthscan: London.

نشریه هوش مصنوعی و داده کاوی

مدل سازی نسبت جذبی سدیم مبتنی بر مدل درختی ام۵ گروهی

محمدتقی ستاری[1*]، ماهش پال[2]، رسول میرعباسی[3] و جان آبراهام[4]

[1*] گروه مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران.

[2] گروه مهندسی عمران، موسسه ملی فن آوری، کروکشترا، ۱۳۶۱۱۹، هاریانا، هند.

[3] گروه مهندسی آب، دانشکده کشاورزی، دانشگاه شهرکرد، شهرکرد، ایران.

[4] گروه مهندسی، دانشگاه ایالتی توماس، امریکا.

ارسال ۲۰۱۷/۰۳/۲۱؛ بازنگری ۲۰۱۷/۰۶/۰۶؛ پذیرش ۲۰۱۷/۰۷/۰۹

چکیده:

در این مقاله چهار رویکرد جمعی برمبنای مدل درختی ام۵ جهت پیش‌بینی نسبت جذبی سدیم (SAR) معرفی گردید. روش‌های جمعی خروجی مدل‌های رگرسیونی چندگانه را با هم ترکیب می‌کنند تا جواب‌های دقیقتری نسبت به زمانی‌که تنها از یک مدل استفاده می‌شود، ارائه دهند. در این مطالعه روش‌های تقویتی فزاینده، کیسه کردن، چرخش جنگل و زیرفضای تصادفی بکار گرفته شد. داده‌های مورد استفاده در این تحقیق مربوط به ۴۴۸ مورد نمونه برداری شامل نه پارامتر از رودخانه باراندوزچای در غرب ایران می‌باشد. سه معیار ارزیابی شامل ضریب همبستگی، ریشه میانگین مربعات خطا و میانگین خطای مطلق برای انتخاب و قضاوت در مورد دقت مدل‌های گروهی استفاده گردید. برای انتخاب متغیرهای ورودی مفید جهت پیش بینی مقادیر SAR دو روش انتخاب ویژگی شامل الگوریتم‌های بسته‌بندی و ژنتیک مورد استفاده قرار گرفت. عملکرد خوب روش‌های مورد استفاده جهت تخمین مقادیر SAR باعث افزایش انگیزه استفاده از این روش‌ها می‌گردد.

کلمات کلیدی: کیفیت آب، نسبت جذبی سدیم، داده‌کاوی، مدل درختی ام۵، الگوریتم ژنتیک، ایران.