Model. Earth Syst. Environ. (2016) 2:105
DOI 10.1007/s40808-016-0165-z

ORIGINAL ARTICLE

Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: a case study of the South Pars Gas Field, Persian Gulf, Iran

Reza Farzi[1] · Vahid Bolandi[2]

Received: 4 May 2016 / Accepted: 4 June 2016 / Published online: 17 June 2016

Abstract Permian–Triassic deposited Dalan–Kangan carbonates in the South Pars Field are the host of the world's largest gas field in which, the Silurian shales might have had triggered the majority of the accumulated gas. Although it is believed that Dalan–Kangan has contribution in gas expulsion. One of the most important factors manipulating the source potential evaluation is considered to be total organic carbon (TOC). Since no TOC no source rock will be found in the area. Consequently in present investigation TOC were utilized as a tool to evaluate the organic facies in combination with the intelligent methods. Current study implements ensemble algorithms as a new method in geoscience data appraisal in comparison with conventional intelligent systems. First of all, we applied fuzzy inference system (FIS) and neural network (NN) as traditional intelligent methods and LSBoost (LSB) and Bagging (BG) as ensemble algorithms to estimating the TOC from well log data. In the next step, Savitzky–Golay filter has been exploited for the data smoothing in order to galvanize regression and classification accuracy. Then, organic facies class membership was taken out by cluster analysis of the synthetized TOC values. In the last place, organic facies class membership were predicted using AdaBoost (AB), LogitBoost (LB), GentleBoost (GB) and Bagging (BG) as ensemble algorithms versus FIS and NN as being conventional intelligent systems directly from petrophysical well log data. Experimental results depict that ensemble methods outperform the common intelligent methods in term of regression and classification concepts. Also, Dalan–Kangan contribution in gas expulsion has been proved by detection of high organic rich interval (parts of k3 unit in Upper Dalan) in the field of study.

Keywords Organic facies · Ensemble algorithm · Cluster analysis · South pars gas field

✉ Reza Farzi
rezafarzi@alumni.ut.ac.ir

Vahid Bolandi
bolandi_v@yahoo.com

[1] School of Geology, University College of Science, University of Tehran, Tehran, Iran

[2] Faculty of Geosciences, Shahid Chamran University, Ahwaz, Iran

Introduction

Quantity, quality and thermal maturity of organic matter are the factors controlling the geochemical evaluating of source rock units. Quantity of organic matter is expressed as total organic carbon (TOC) (Hunt 1996), quality is controlled by type of kerogen on the other hand Hydrogen Index (HI) and thermal maturity defined by T_{max} parameter taken out by Rock–Eval analysis. As stated by Jones (Jones 1987) organic facies based on geochemical parameters is in concordance with depositional facies belts, wherever circumstances of each individual subenvironment gives rise to similarity in geochemical features.

As a result of the uncertainty associated with data and non-linear relation between the geosciences data using the unconventional mathematical methods are inevitable (Nikravesh et al. 2003). So in recent decades advent of the machine learning played a key role in earth science. In this concern, fuzzy logic, neural networks, and genetic algorithms are the most chosen machine learning techniques used to solve the modeling problems (Saggaf and Nebrija 2003; Kadkhodaie-Ilkhchi et al. 2006). While ensemble methods which are known as powerful supervised classification and regression techniques scarcely has been applied in earth science data analysis (Kadkhodaie-Ilkhchi et al. 2010; Monteiro et al. 2009). Indeed ensemble learning is a powerful technique for combining multiple base classifiers or predictor to produce a form of committee whose performance can be significantly better than that of any of the base classifiers or predictor (Bishop 2006; Rokach 2010). So in current study, ensemble regression methods (LSBoost and Bagging) and ensemble classification techniques (AdaBoost, LogitBoost, GentleBoost and Bagging) were used to predict TOC and organic facies class membership (obtained by cluster analysis of the predicted TOC values) respectively in comparison with conventional intelligent systems (fuzzy logic and neural network). Moreover, Savitzky–Golay smoothing filter is applied as a pre-processing step to attain a good performance in term of regression and classification accuracy.

The abovementioned methods are applied in The South Pars gas field which is known as one of the largest non-associated gas field in the world being located in the Persian Gulf (Ghazban 2007) (Fig. 1), in which gas accumulation is mostly limited to the Permian–Triassic stratigraphic units known as Dalan–Kangan Formations composed of carbonate-evaporate series (Aali et al. 2006) (Fig. 2). The source of this accumulation is considered to be triggered off from the early Silurian shales as stated by Aali et al. (2006). While, contribution of Permian–Triassic units in production had been fully discussed in Galimov and Rabbani (2001). So we decided to resolve the contradiction by joint use of the Rock–Eval parameters and organic facies of the Dalan–Kangan carbonates for appraisal of the Dalan–Kangan production contribution.

Methods and materials

In recent decade, essential to reduce uncertainty and time and cost consumption in geoscience investigations inclined researchers to consider application of machine learning methods in data analysis (Kadkhodaie-Ilkhchi et al. 2006, 2010; Monteiro et al. 2009).

Machine learning is a subfield of computer science and so-called artificial intelligence that deals with the construction and study of systems that can learn from data, rather than following only explicitly programmed instructions. Among these techniques ensemble learning refers to the procedures employed to train multiple learning machines and combine their outputs, treating them as a committee of decision makers (Brown 2010). The principle is that the committee decision, with individual predictions combined appropriately, should have better overall accuracy, on average, than any individual committee member (Brown 2010; Bishop 2006). To put it simply, an ensemble is a technique for combining many weak learners in an attempt to produce a strong learner.

In this study, machine learning techniques are used for supervised learning of classifiers and predictors. With the aim of comparing between ensemble algorithms and conventional intelligent methods (Fuzzy Logic and Neural Networks), Regression ensemble model (LSBoost and Bagging) were constructed to model TOC values from well log data. In the next step, K-means clustering applied to synthetize organic facies class membership needed in

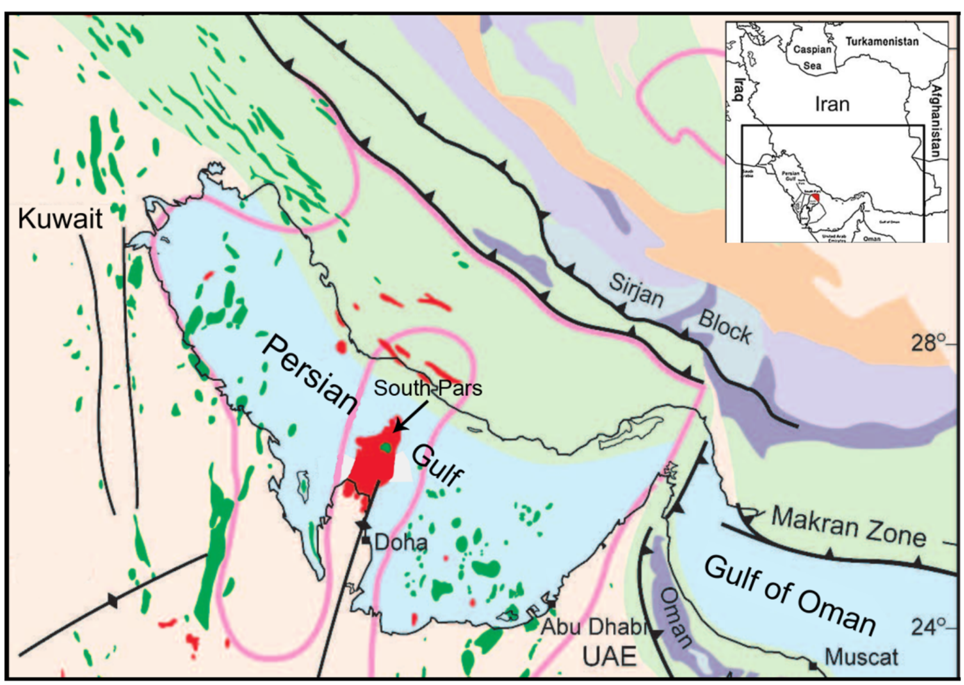

Fig. 1 Geographical and geological setting of the South Pars gas field. Main hydrocarbon fields in Persian Gulf and adjacent areas are shown (Modified from Insalaco et al. 2006)

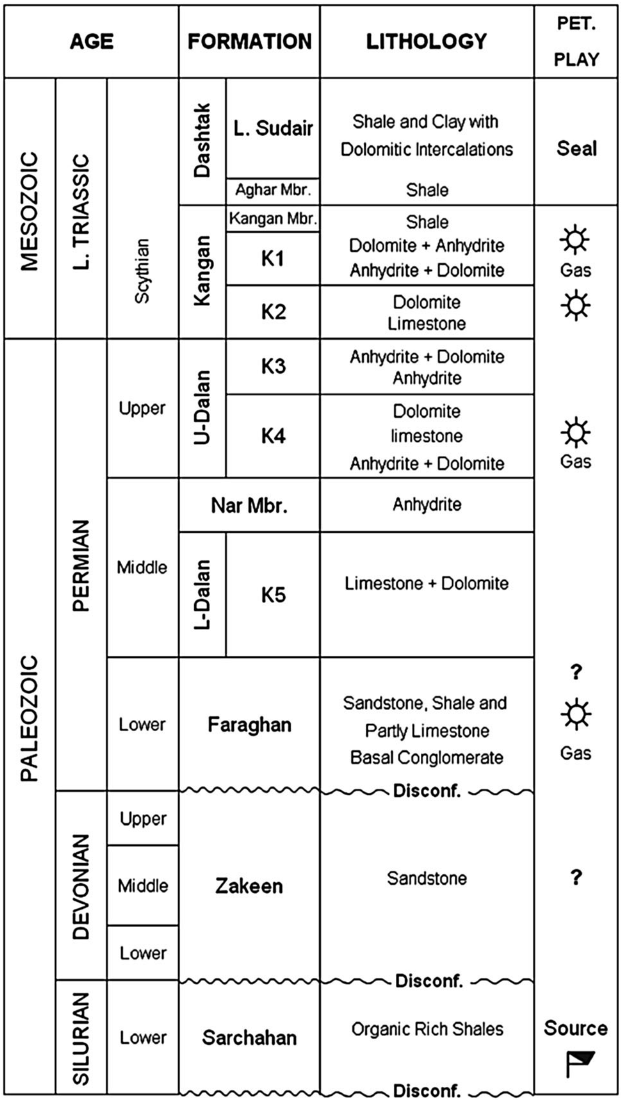

Fig. 2 Stratigraphic chart for the South Pars field (Aali et al. 2006)

term of classification concepts. After that, Classification ensemble engines (AdaBoost, LogitBoost, GentleBoost and Bagging) architected for identification of organic facies class membership from well log data and results compared against common intelligent classifiers performance.

Boosting

Boosting is a powerful method for improving the performance of any learning algorithm. It can be used to significantly reduce the error of any "base" learning algorithm that consistently generates classifier which need only be a little bit better than random guessing (Freund and Schapire 1996). Boosting has several kinds, among those variants; we are particularly focusing on AdaBoost, LogitBoost, GentleBoost and LSBoost. At each stage of the algorithm, training sets and classifiers generated sequentially based on the results of the previous iteration (Friedman et al. 2000) which weighting coefficients of misclassified data are greater. According to Fig. 3a, each base classifier $f_i(x)$ is trained on a weighted form of the training set (blue arrows) in which the weights w_i depend on the performance of the previous base classifier $f_{i-1}(x)$ (green arrows). Once all the base classifiers have been trained, they are combined to give the final classifier $F_M(x)$ (red arrows). We denote a training dataset by $\{x_i, y_i\}_{i=1}^{N}$, where N is the number of samples, x_i is the ith feature vector, and $y_i \in \{0, 1, 2, \ldots, K-1\}$ is the ith class label, where $K \geq 3$ in multi-class classification. Here the inputs are conventional well log data and organic facies class membership is the output value. Several Boosting Algorithms are as follows:

AdaBoost

The most commonly used version of Boosting is AdaBoost (Adaptive Boosting), in which stage-wise gradient descent procedure is used in an exponential cost function (Friedman et al. 2000). One of the main ideas of the algorithm is to maintain a distribution or set weights over the training set (Schapire 1999) thereby after each iteration of algorithm the weights of misclassified data point are increased. AdaBoost is a binary classifier and can be extended to handle multiclass problems using one-versus-all strategy. The formal algorithm is:

(a) start with weights $w_i = 1/N$, $i = 1, 2, \ldots, N$.

(b) repeat for $m = 1, 2, \ldots, M$:

   (i) fit the classifier to obtain a class probability estimate $p_m(x) = \hat{P}_w(y = 1|x) \in [0, 1]$ using weights w_i on the training data

   (ii) set $f_m(x) \leftarrow \frac{1}{2} \log p_m(x)/(1 - p_m(x)) \in R$

   (iii) set $w_i \leftarrow w_i \exp[-y_i f_m(x_i)]$, i = 1, 2, …, N, and renormalize so that $\sum_i w_i = 1$

(c) output the classifier $sign[\sum_m^M f_m(x)]$

GentleBoost

GentleBoost is another version of AdaBoost named Gentle AdaBoost as well, which outperforms in proportion to prior version (Friedman et al. 2000). GentleBoost uses adaptive Newton steps to optimize the cost function of the classifier rather than exact optimization at each step. The main

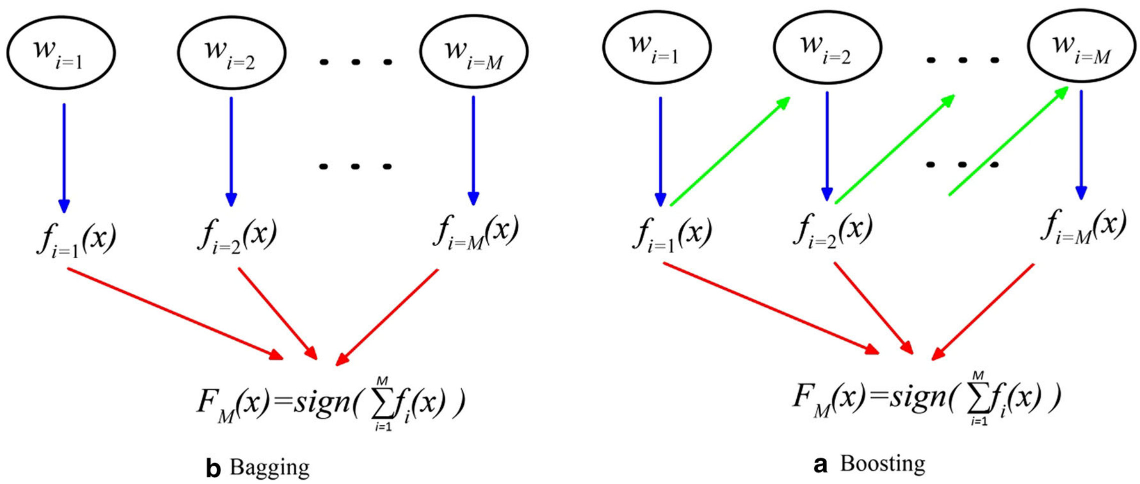

Fig. 3 Schematic illustration of the boosting and bagging framework (Modified from Bishop 2006)

difference between GentleBoost and AdaBoost is how it uses its estimates of the weighted class probabilities to update the functions (Friedman et al. 2000). In this case one-versus-all method was used to tackle multiclass issue. The formal algorithm is:

(a) start with weights $w_i = 1/N$, $i = 1, 2, \ldots, N$, $F(x) = 0$.

(b) repeat for $m = 1, 2, \ldots, M$:

  (i) fit the function $f_m(x)$ by weighted least-squares of y_i to x_i with weights $w_{i.}$

  (ii) update $F(x) \leftarrow F(x) + f_m(x)$

  (iii) update $w_i \leftarrow w_i \exp(-y_i f_m(x_i))$ and renormalize.

(c) output the classifier

$$sign[F(x)] = sign[\sum_{m=1}^{M} f_m(x)] \quad (1)$$

LogitBoost

The LogitBoost is a variant of Boosting, capable of handling multiclass problems, through which, stagewise optimization of the maximum likelihood has been used within adaptive Newton steps to fit additive logistic regression models (Friedman et al. 2000). In order to tackle the multiclass problems it utilizes a symmetric multiple logistic transformation

$$pj(x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}} \quad (2)$$

$$\sum_{k=1}^{J} F_k(x) = 0 \quad (3)$$

where p_j is the probability of assigning class c among C classes. The formal algorithm is:

(a) start with weight $w_{ij} = 1/N$, $i = 1, \ldots, N$, $j = 1, \ldots, J$, $F_j(x) = 0$ and $p_j(x) = 1/J \forall j$.

(b) repeat for $m = 1, 2, \ldots, M$:

  (i) repeat for $j = 1, \ldots, J$:

    * Compute working response and weight in the jth class,

$$z_{ij} = \frac{y_{ij} - p_j(x_i)}{p_j(x_i)(1 - p_j(x_i))} \quad (4)$$

$$w_{ij} = p_j(x_i)(1 - p(x_i)) \quad (5)$$

    * fit the function $f_{mj}(x)$ by a weighted least-squares regression of z_{ij} to xi with weight w_{ij}

  (ii) set $f_{mj}(x) \leftarrow \frac{J-1}{J}(f_{mj}(x) - \frac{1}{J}\sum_{k=1}^{J} f_{mk}(x))$ and $F_j(x) \leftarrow F_j(x) + f_{mj}(x)$.

  (iii) update pj(x) via,

$$pj(x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}, \sum_{k=1}^{J} F_k(x) = 0 \quad (6)$$

(c) output the classifier arg max$_j$ $F_j(x)$.

LSBoost

The LSBoost algorithm fits regression ensembles that at every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown previously (Hastie et al. 2008). The ensemble fits to minimize mean-squared error. The formal algorithm is:

(a) start with $i = 1, 2, \ldots, N$, $F(x) = 0$.

(b) repeat for $m = 1, 2, \ldots, M$:

(i) fit the regression function $f_m(x_i)$ to x_i.
(ii) update $F(x) \leftarrow F(x) - \eta f_m(x_i)$.

(c) output the predictor $\frac{\sum_{m=1}^{M} F(x)}{N}$.

The algorithm uses η as shrinkage by passing in learning rate parameter that the range of this parameter is 0 to 1.

Bagging

The so called Bagging algorithm (Bootstrap aggregating) was proposed by Breiman in 1994 to improve the learning by combining learners of randomly generated training sets (Breiman 1996) (Fig. 3b). Bagging predictors is a method for generating multiple version of a predictor and using these to achieve an aggregated predictor (Breiman 1996). The reason behind Bootstrap aggregating development is to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression techniques. Moreover, this technique reduces variance and helps to avoid over-fitting as opposed to Boosting. As stated the Bagging is a special case of the model averaging approach and most effective one when the base learner is unstable (Brazdil et al. 2009). The formal algorithm is as follow:

(a) For $k = 1$ to N

(i) S_k = random sample of size d drawn from T, with replacement
(ii) h_k = model induced by A from S_k

(b) For each new query instance q
(c) For classification:

$$\text{Class}(q) = \arg\max_{y\in\gamma} \sum_{k=1}^{N} \delta(y, h_i(q)) \quad (7)$$

for regression:

$$\text{Value}(q) = \frac{\sum_{i=1}^{N} h_i(q)}{N} \quad (8)$$

where T is the training set, A is the chosen learning algorithm, N is the number of samples or bags, each of size d, drawn from T, γ is the finite set of target class values, δ is the generalized Kronecker function ($\delta(a,b) = 1$ if $a = b$; 0 otherwise)

Fuzzy systems

A fuzzy expert system is a kind of system that utilizes fuzzy logic that statements are no longer black or white, true or false, on or off versus Boolean logic that takes on a value of either zero or one (Sumathi and Surekha 2010). Therefore, FL is well suited to solve geosciences problem which are associated with uncertainty and vagueness (Kadkhodaie-Ilkhchi et al. 2006). Fuzzy inference system (FIS) is the procedure of formulating between input and output using fuzzy logic, first introduced by Zadeh (1965). The formulation between inputs and outputs consist of if–then rules which extracted through a fuzzy clustering process (Kadkhodaie-Ilkhchi et al. 2010), where subtractive clustering is the most widely used approach for effective construction of a fuzzy model through search for the optimal clustering radius, which is a controlling parameter for determining the number of fuzzy if–then rules.

Neural networks

Generally, neural networks can be considered as an information processing system composed of neurons which are networks of interconnected simple processing elements capable of learning any input–output relationship. Feed-forward network with back-propagation learning rule considered one of the efficient types of the neural networks (Bishop 2006). The back-propagation rule is comprised of two paths (Kadkhodaie-Ilkhchi et al. 2010):

(a) Forward path: input vector conducts on feed-forward network thus obtained values propagate to output layer through hidden layer.

$$Y_s = g(\sum_j \beta_{sj}(t)(g \sum_i a_{ji}(t) I_i(t))) \quad (9)$$

where I is input vector, Y is network output, α and β are the first and second layer weights, g is the activation function and t is forward and backward times.

(b) Backward path: during this step, the parameters of network change and adjust which this adjustment conduct based on error-correcting learning rule; specifically a sum of squares error measure (E) is calculated.

$$E(t) = \frac{1}{2}\sum_{s=1} (d_s(t) - Y_s(t))^2 \quad (10)$$

where d is the target value for dimension s.

K-Means clustering

Cluster analysis is a popular unsupervised categorizing technique used to discover uncovered relationships within data. There are some clustering techniques available that the K-means method is one of the popular algorithms amongst, first introduced by MacQueen (1967). K-means is an algorithm to partitioning data into k mutually exclusive clusters and assign a specific number of centers, k, to represent the clustering of N points ($k > N$) (Nikravesh

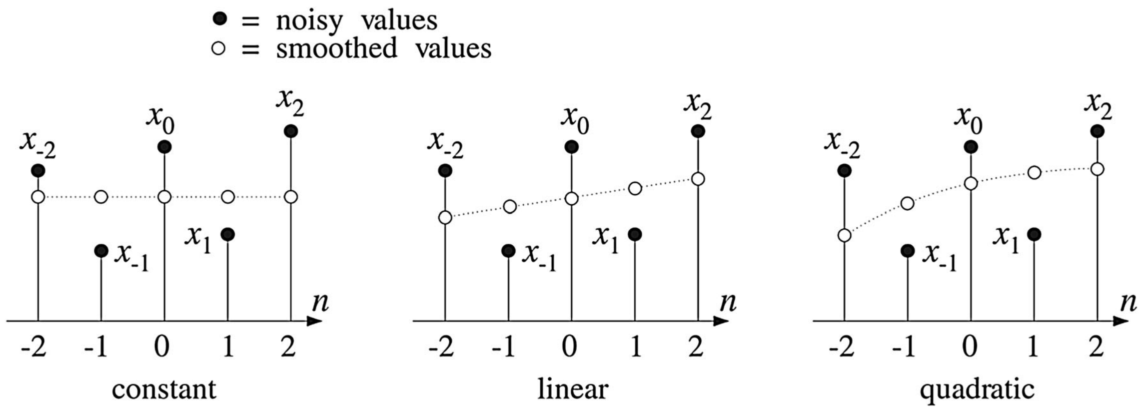

Fig. 4 Data smoothing with polynomial of degree d = 0, 1, 2 (Orfanidis 2010)

et al. 2003). K-means uses an iterative algorithm that minimizes the sum of distances from each point to its cluster centroid, over all clusters. In general, the k-means technique generates exactly k different clusters of the greatest possible distinction.

The algorithm is summarized as following (Nikravesh et al. 2003):

(a) consider each cluster consisting of a set of M samples that are similar to each other: $x_1, x_2, x_3, \ldots, x_m$.
(b) choose a set of clusters $\{y_1, y_2, y_3, \ldots, y_k\}$.
(c) assign the M samples to the clusters using the minimum Euclidean distance rule.
(d) compute a new cluster so as to minimize the cost function.
(e) if any cluster change, return to step c; otherwise stop.
(f) end.

The Silhouette validation method

The fundamental of Silhouette index is defining the mean distance between each point and cluster center. For each point silhouettes width $s(i)$ are formulating as follow and ranges from −1 to 1 (Tan et al. 2006):

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), B(i))} \tag{11}$$

where $a(i)$ is mean distance of i-point to all other point in the same cluster; $b(i)$ is the minimum of mean distance of i-point to all point in other cluster.

Silhouettes width is used to validating performance of given clustering method as when $s(i)$ close to 1 the sample is well-clustered. Moreover, it is utilized to identifying the optimal number of clusters (Sfidari et al. 2012). The mean of the silhouette width for a given cluster C_k (cluster mean silhouette) is denoted as s_k and used for determination of the optimum number of the:

$$s_k = \frac{1}{n_k} \sum_{i \in I_k} s(i) \tag{12}$$

Savitzky–Golay filtering

The former studies confirm that smoothing analysis increases the accuracy of classification and regression of intelligent engines (Monteiro et al. 2009). In other words, in attempting to smooth out the noise, the filter begins to smooth out the fairly narrow peaks of the desired signal (Orfanidis 2010). The Savitzky–Golay filters (SGF), also known as polynomial smoothing, or least-squares smoothing filters are widely used for smoothing and differentiation in many topics (Bromba and Ziegler 1979; Ziegler 1981). Savitzky–Golay filters minimize the least-squares error in fitting a polynomial to frames of noisy data (Savitzky and Golay 1964) (Fig. 4).

Prediction of TOC

One of the most important factors manipulating source rock evaluation is TOC, which is best derived by Rock–Eval pyrolysis. With the goal of gaining TOC values, 74 cutting samples obtained from two wells in the South Pars Field exposed to Rock–Eval pyrolysis. Due to lack of geochemical data from entire sequence attempt has been made to model this parameter from well log data as a common data which is available in almost all wells. For this reason well log data including DT, NPHI, RHOB, GR, PEF and LLD has been prepared to model TOC values for the entire sequence. Prior to designing an intelligent model appropriate connection between inputs and outputs should be established. Consequently, the good relationship among TOC and petrophysical data can be seen in Fig. 5. At this point, the dataset including TOC values and their

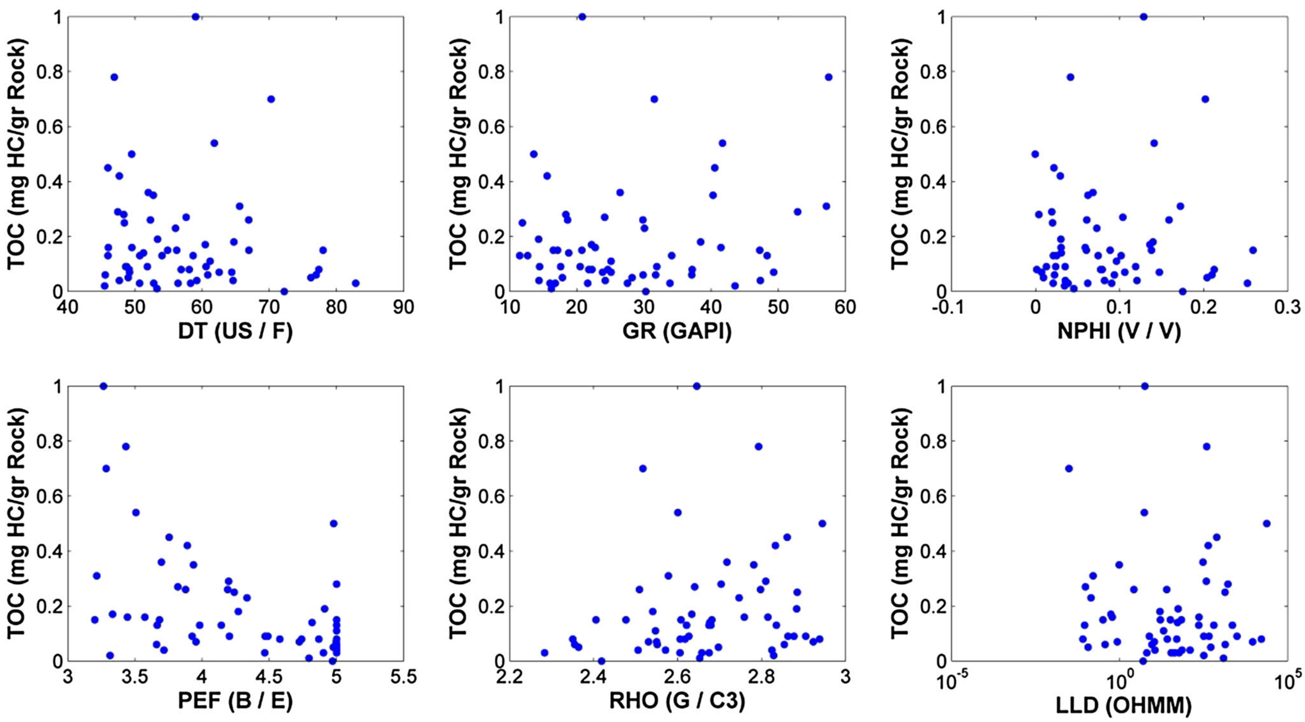

Fig. 5 Cross-plot showing relationship between real TOC and well log data including DT (**a**), GR (**b**), NPHI (**c**), PEF (**d**), RHOB (**e**), and RT (**f**) in training data

corresponding well log data were divided into 60 and 14 training and test sets, respectively. Test sets (cross validation) taken out to assess the reliability of the constructed models. In the following parts attempt to characterize the TOC values using intelligent systems are fully discussed.

In this step, a Takagi–Sugeno fuzzy system (TS-FIS) and Feed-Forward neural network (FF-NN) were selected to predicting TOC from well log data as being successful intelligent methods in machine learning fields. In the Following step, results were compared to ensemble method and conclusion has been drawn. Using MATLAB® software, a Takagi–Sugeno model was created to synthesis TOC values from well log data. In this case, fuzzy if–then rules, inputs and output membership functions were extracted by subtractive clustering method as powerful and efficient method for fuzzy logic modeling (Chiu 1994; Kadkhodaie-Ilkhchi et al. 2010). Since cluster radius which controls the cluster numbers and fuzzy rules is significant parameter in fuzzy logic modeling, clustering process was accomplished gradually with cluster radius from 0.005 to 1 (with 0.005 intervals) to find optimal cluster radius. Consequently, minimum root mean square error (RMSE) corresponds to cluster radius 0.66 resulting in 4 fuzzy if–then rules for a model with 6 inputs and 1 output (Fig. 6a). In the constructed model four Gaussian membership functions fitted to the derived input clusters and output membership function is linear type equation. In order to formulate input petrophysical data to TOC values a three layer Feed-Forward neural network model constructed. In view of most prior researches it is confirmed that a three layered neural network outperforms those with more layers in term of processing time and accuracy (Balkin and Ord 2000; Bolandi et al. 2015; Khan et al. 2008; Kadkhodaie-Ilkhchi et al. 2006, 2008, 2010). The essential parameters required in generalization of the neural network models are number of the hidden layer, number of the neurons in the hidden layer, the transfer functions, the training function and the training epochs. The essential parameters of the established network were defined as below:

Input layer composed of six neurons based on six inputs and the number of neurons in one hidden layer by implementing a gradually process against RMSE has been defined as 4 (Fig. 6b). The Levenberg–Marquardt algorithm was found to be the proper training function for regulating the weights and biases according to the error. The transfer function between input layer and hidden layer is tangent logistic and from hidden layer to output layer is linear. The network succeeds to the minority of MSE in 23 epochs.

In this part, applicability of two variants of ensemble algorithms including LSBoost and regression version of Bagging is being investigated in term of TOC prediction from conventional well log data, both of which using regression tree as weak learner. One of the major steps in

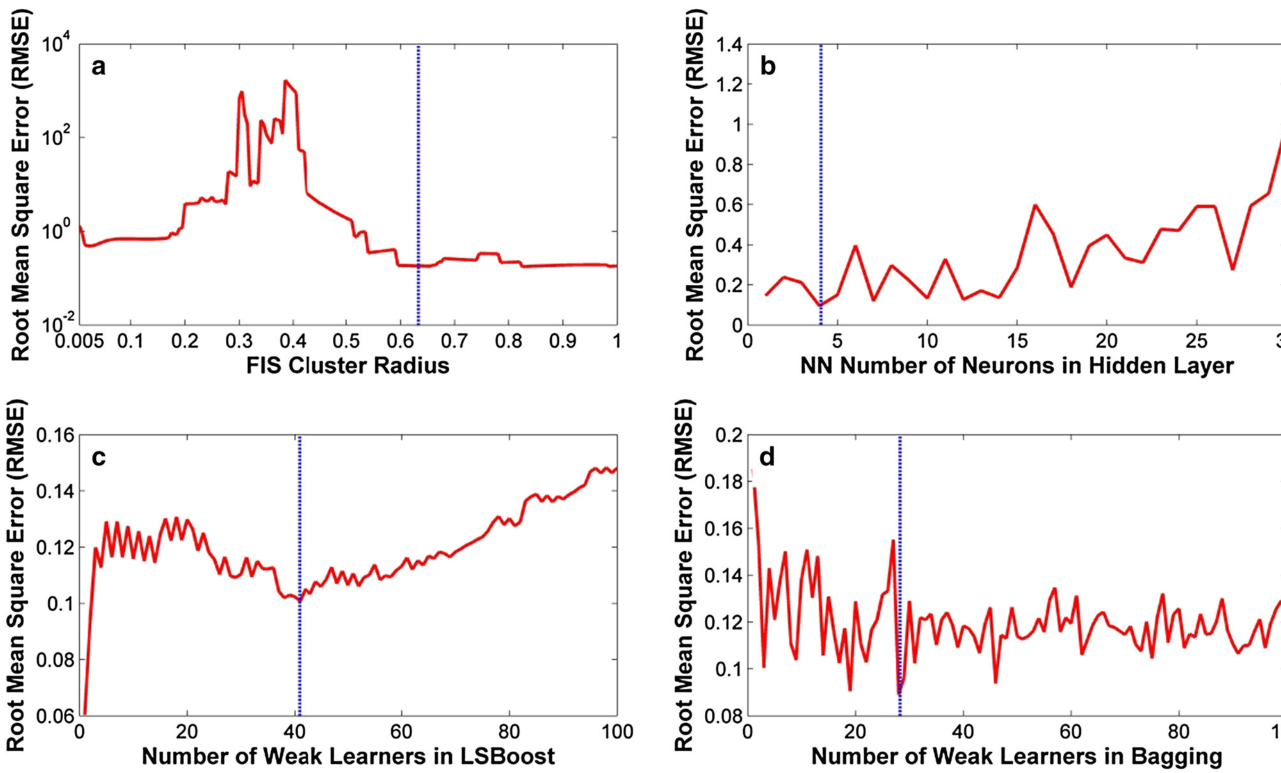

Fig. 6 RMSE versus FIS clustering radius (**a**) NN number of neurons in hidden layer (**b**) number of weak learner in LSBoost (**c**) and Bagging (**d**). *Blue line* shows optimal value

exploiting these algorithms is to specify the number of weak learners which is the only predefined parameter needed, stated as privilege of ensemble methods. Choosing the optimal size of an ensemble involves balancing between speed and accuracy so that larger ensembles take longer time to train and some ensemble algorithms can become overtfitted in case of too large values. Therefore regression process was accomplished gradually from 1 to 100 (with 1 intervals) to achieving optimal number of weak learners. Then, the models with the highest overall accuracy were selected as the optimal models based on the RMSE of trained models versus number of weak learners (Fig. 6c, d). The results show that taking the number of weak learners to 41 and 29 leads to the highest performance in LSBoost and Bagging algorithms, respectively.

Specification of organic facies

Generally speaking, Jones (1987) defined an organic facies as a mappable subdivision of a stratigraphic unit distinguished by the character of its organic matter, whereas we specify it by clustering quantity of the organic matter within the sequence of study. Since cluster analysis is powerful method for uncovering relationships in large multivariate data sets, therefore K-means technique was applied to generate organic facies class membership from synthetized TOC values prior to classification task in which each input value assigned to one of a given set of classes (organic facies class memberships). We tackle the choice of optimal cluster numbers, by taking into account as main issue in applying cluster algorithms, utilizing Silhouette value as validity measurement. The results show that taking the number of clusters to 3 leads in best data labeling (Fig. 7).

Estimation of organic facies

In the previous step cluster analysis has been discussed that is among exploratory data analysis techniques. These techniques attempt to analyze data without directly using information about the class assignment in this case organic facies class membership of the samples. Even so cluster analysis is known as powerful methods for revealing uncovering relationships in large multivariate data sets, they are not sufficient for developing classification rules that can accurately predict the class-membership of unknown samples. Consequently, ensemble approaches and conventional intelligent methods were applied for prediction of the organic facies class membership from conventional well log data instantly, which is the main

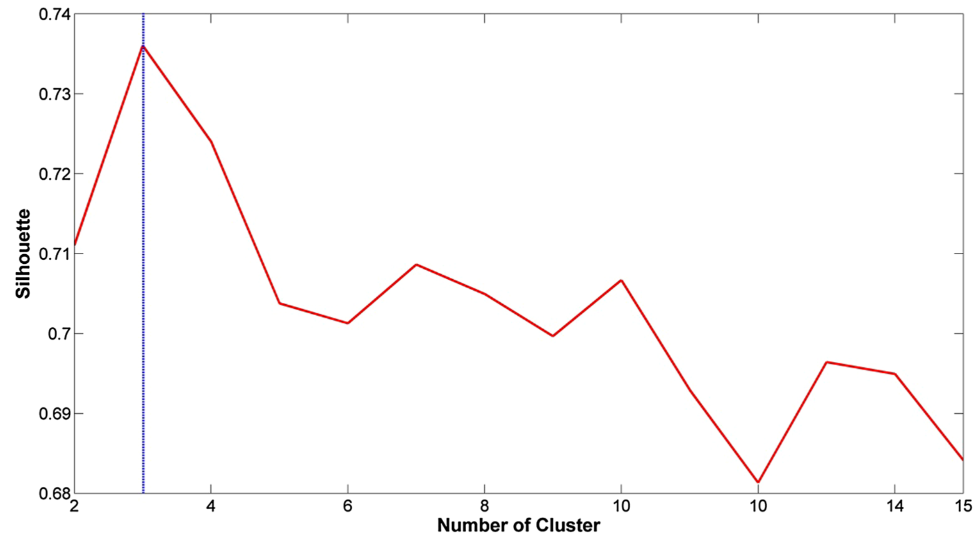

Fig. 7 Silhouette value of clusters by K-means clustering approach to selecting best cluster number

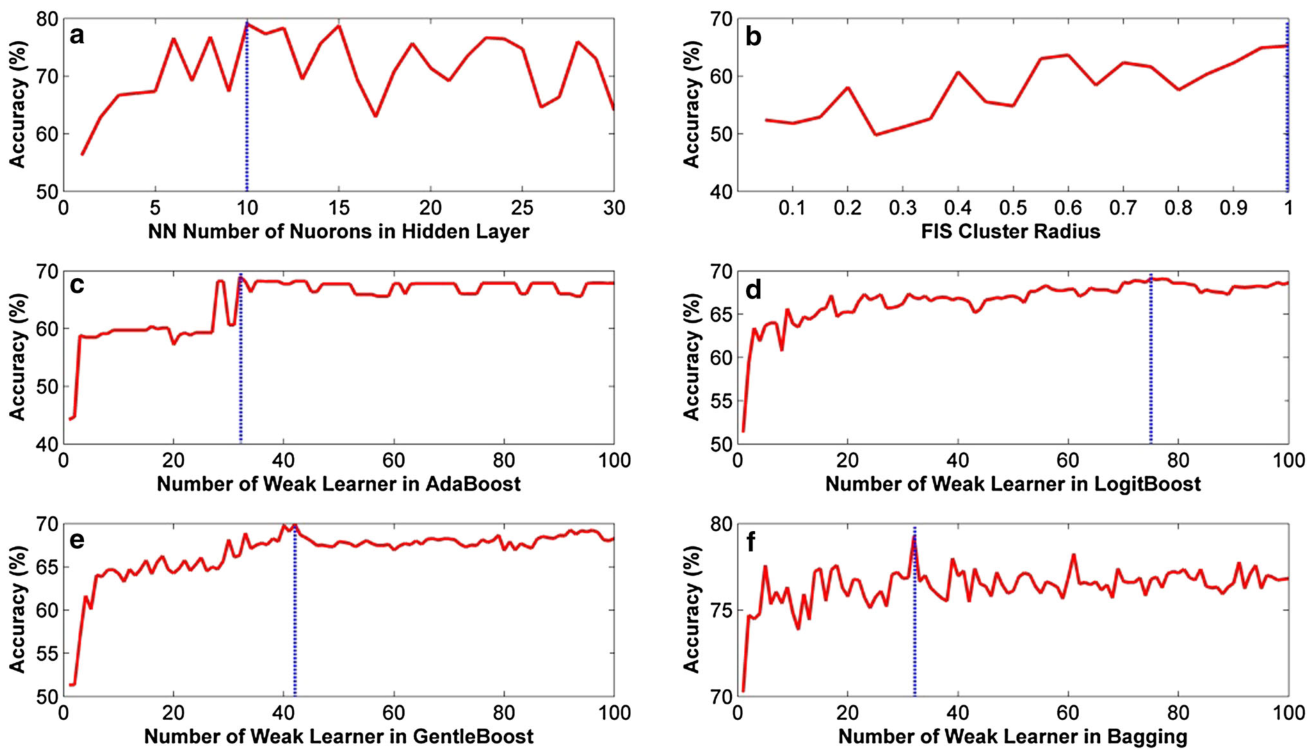

Fig. 8 Accuracy versus FIS clustering radius (**a**) NN number of neurons in hidden layer (**b**) and number of weak learner in AB (**c**), LB (**d**), GB (**e**), BG (**f**). *Blue line* shows optimal value

objective of this paper. Finally, results have been compared and conclusion has been drawn. Within this step corresponding data to well SP-A1 and SP-A2 is defined as training and test data sets, respectively.

In this part, ensemble approaches including AdaBoost, LogitBoost, GentleBoost and Bagging in comparison with conventional intelligent methods covering neural network and fuzzy logic algorithms were applied for prediction of the organic facies class membership from conventional well log data instantly. To begin with, a three-layered Feed-Forward network was designed using MATLAB software and trained with ten neurons in the hidden layer (Fig. 8a). In the following step, comparative study performed by applying fuzzy logic classifier as a class-membership predictive model and

Table 1 Performance of machine learning methods in predicting TOC and effect of Savitzky–Golay smoothing filter

Methods	RMSE without SGF	RMSE with SGF
BG	**0.1128**	**0.0895**
LSB	0.1964	0.1002
NN	0.1845	0.0941
FIS	**0.2173**	**0.1304**

The bold values represent the best and worst performance between the methods

as a matter of course, the cluster radius were determined to be 1 (Fig. 8b). Then, four types of ensemble methods, AdaBoost, LogitBoost, GentleBoost and Bagging were inspected to classify the petrophysical well log data into organic facies. Regardless of little disparity of these algorithms in exerting the weights on each weak learner, determination of the number of weak learners is the fundamental parameter manipulating the performance of the ensemble algorithms. Consequently, training process was accomplished gradually to achieving optimal number of the weak learners from 1 to 100. As a result, the optimal number of the weak learners detected to be 32, 42, 75 and 32 with relevant maximum accuracy for AdaBoost, GentleBoost, LogitBoost and Bagging, respectively (Fig. 8c–f).

Results and discussion

According to Table 1 the proposed methods is found effective in predicting TOC values from well log data, while Bagging outperforms the others. Quantitative comparisons of the results represents that Bagging algorithm readily outperforms neural network, fuzzy logic and LSBoost in term of accuracy (Figs. 9, 10). Further investigation showed that utilizing Savitzky–Golay smoothing filter as a pre-processing technique boosts the accuracy of models. Beside, high RMSE value on FIS is indicative of the high sensitivity to the data type (Kadkhodaie-Ilkhchi et al. 2010). Also results show that within LSBoost algorithm by increasing number of weak learners up to 50, RMSE decreases while number of weak learners increasing from 50 to 100 leads to increasing in RMSE. This is the case which is known to be address as overfitting. Nevertheless, approximately constant error of Bagging algorithm shows its resistance against overtraining, a noticeable advantage of this algorithms in comparison with LSBoost algorithm. Generated organic facies by k-means clustering and TOC values versus depth in well SP-A1 and SP-A2 are visually illustrated in Fig. 11.

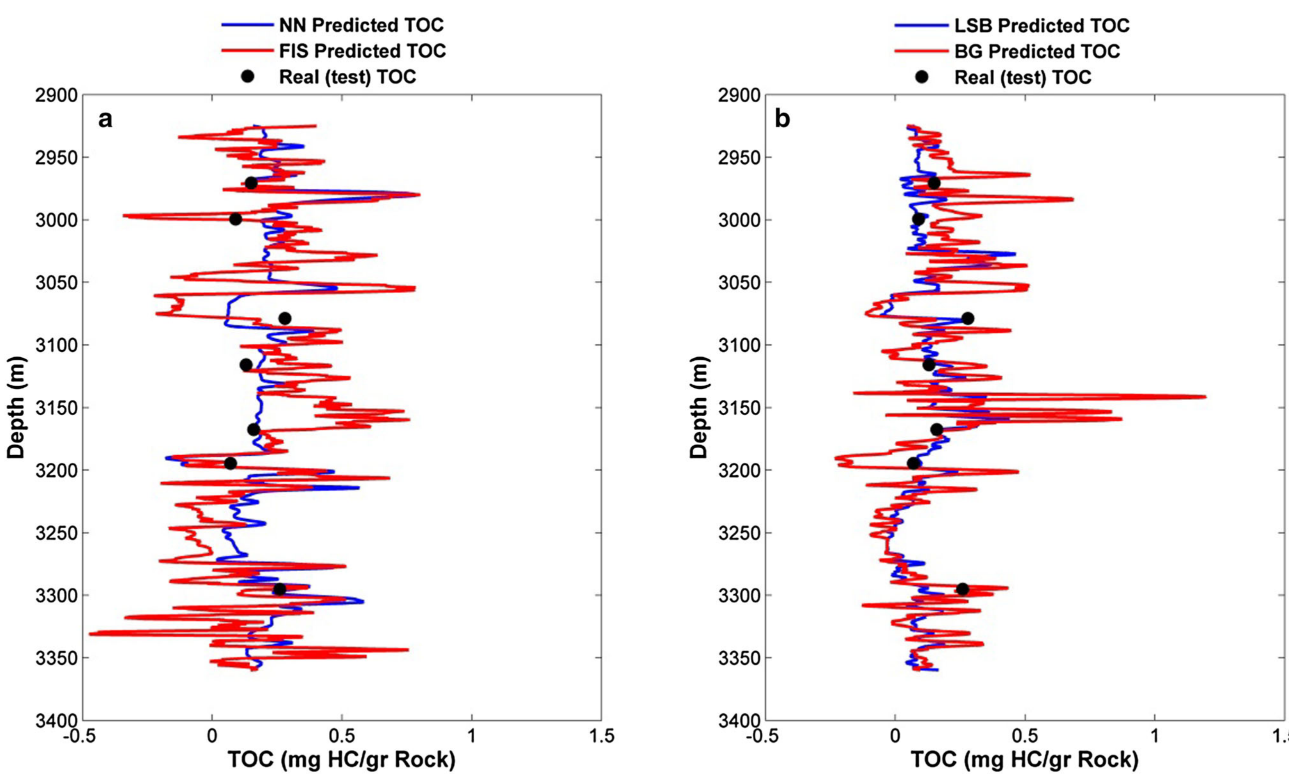

Fig. 9 Comparison of real (test) TOC and predicted TOC using NN and FIS (**a**), BG and LSB (**b**) for well SP-A1

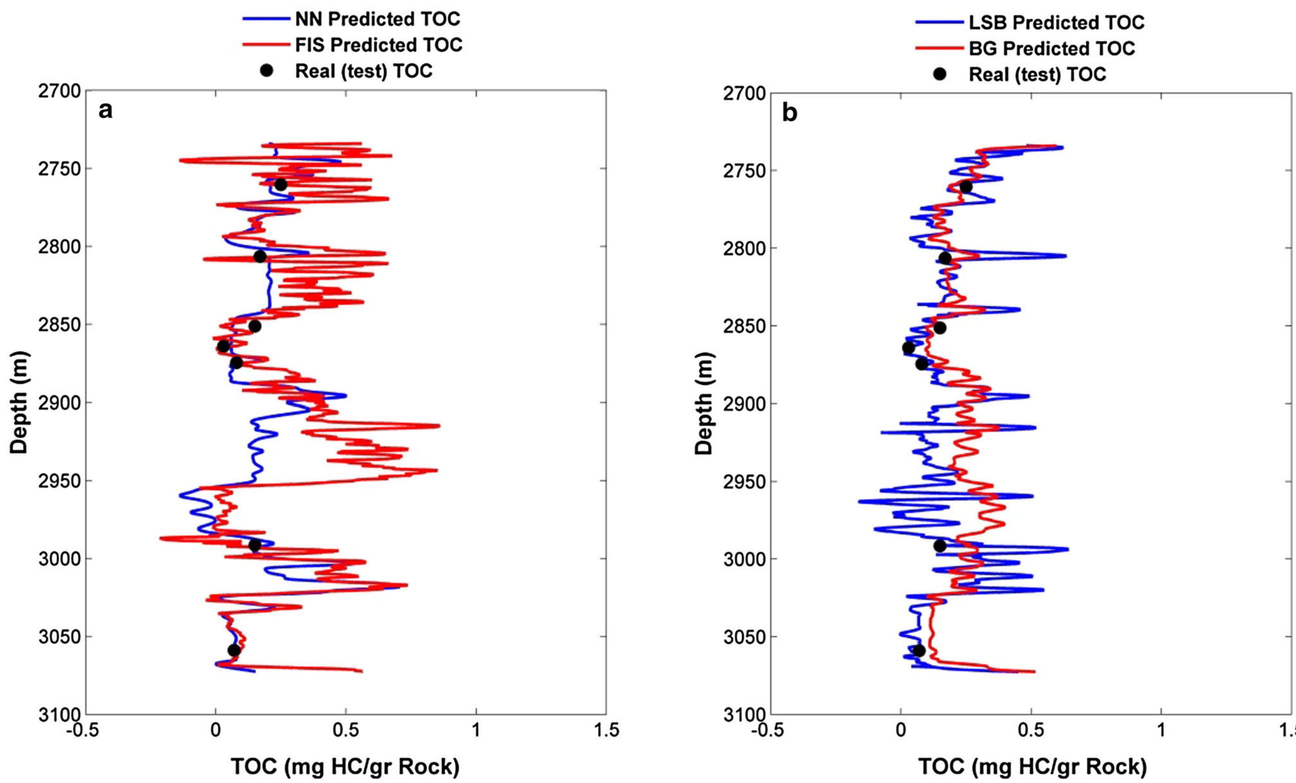

Fig. 10 Comparison of real (test) TOC and predicted TOC using NN and FIS (**a**), BG and LSB (**b**) for well SP-A2

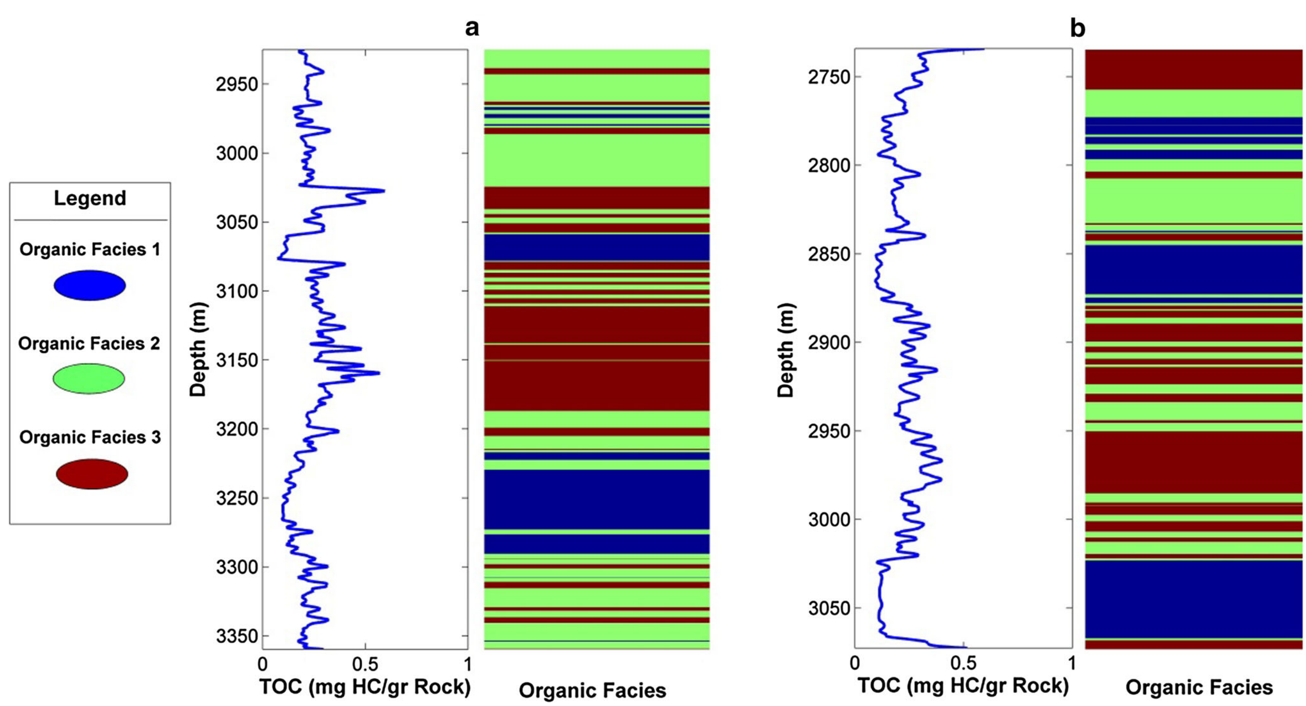

Fig. 11 Created organic facies using K-means clustering for well SP-A1 (**a**) and SP-A2 (**b**)

On the whole, machine learning methods are found effective in specifying the organic facies from well log data instantly. Unlike the normalization technique that has shown no influence on the performance of ensemble methods (Kadkhodaie-Ilkhchi et al. 2010), utilizing SGF as a preprocessing phase on the data substantially boosts the accuracy. Quantitative comparisons of the results depicts that for classification problem of organic facies specification Bagging readily outperforms other techniques in term of classification accuracy, whereas fuzzy system has the lowest performance amongst (Table 2). Moreover, based on the results AdaBoost, GentleBoost and LogitBoost have shown almost similar performances and accuracy of the neural network model is roughly near to Bagging. Comparison of predicted organic facies with real ones demonstrated in Fig. 12. To sum up as can be seen in Fig. 8c–f, resistance of ensemble methods to overtraining can be substantiated by constant value of accuracy for number of weak learners above 20.

Table 2 Performance of machine learning methods in estimating organic facies and effect of Savitzky–Golay smoothing filter

Methods	Accuracy without SGF	Accuracy with SGF
AB	59.72	69.03
GB	63.5	69.89
LSB	63.99	69.12
BG	**68.09**	**79.26**
NN	66.92	79.02
FIS	**58.19**	**65.21**

The bold values represent the best and worst performance between the methods

Conclusion

The main objective of this investigation was to assess the applicability of the ensemble approaches as a new machine learning methods in analysis of intricate geological data which are found effective in specifying the organic facies from well log data instantly. Consequently, TOC values synthetized for entire sequence of study from conventional well log data. Then, synthetized TOC values grouped optimally using K-means cluster analysis as main parameter required for construction of classification rules using ensemble and conventional methods, which leads in 3 organic facies class membership. The main conclusion to be drawn from this discussion is that the ensemble algorithms in particular Boosting and Bagging algorithms outperform conventional machine learning methods including fuzzy logic and neural networks in term of regression and classification concepts. Moreover, it has been concluded that preprocessing data using the Savitzky–Golay smoothing filter results in substantial accuracy boost, while other preprocessing techniques like normalizing show little or no manipulation on performance of models.

Further to abovementioned conclusion, detection of high organic rich interval (parts of k3 unit in Upper Dalan) in the field of study proves Permian–Triassic sequences (Dalan–Kangan) contribution in gas expulsion, which was

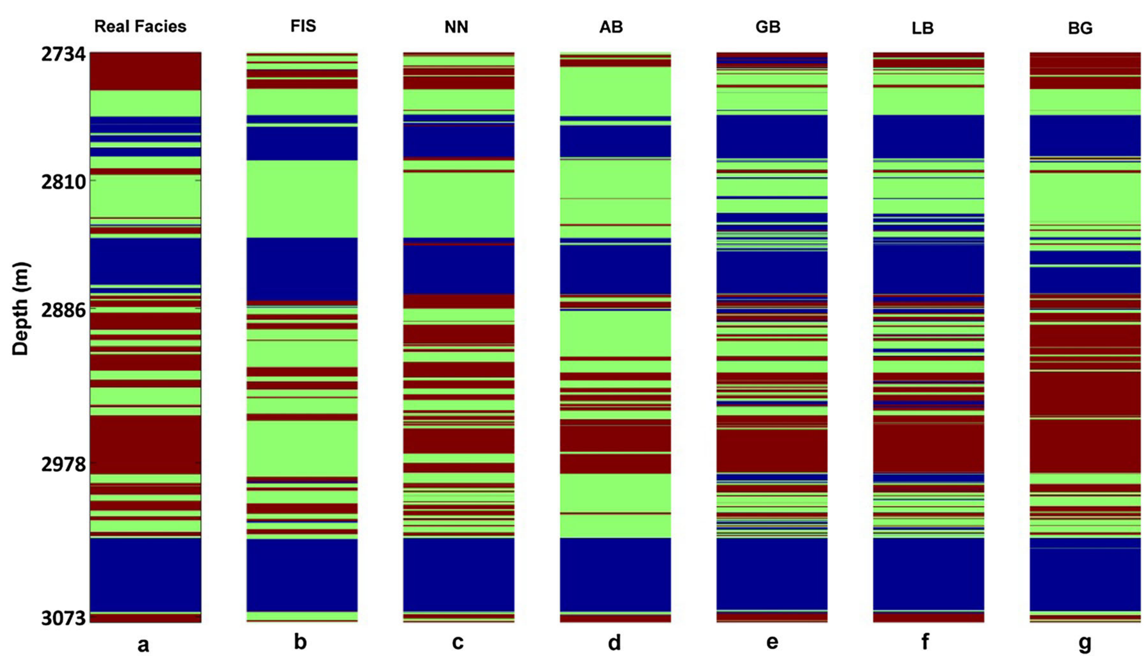

Fig. 12 Comparison of real organic facies (**a**) and estimated using FIS (**b**), NN (**c**), AB (**d**), GB (**e**), LSB (**f**), BG (**g**) in well SP-A2

a contradictory issue whether it has been contributed in gas expulsion or not.

Acknowledgments The authors appreciate Pars Oil and Gas Company (POGC) of Iran for financial support, data preparation and permission to publish this paper.

References

Aali J, Rahimpour-Bonab H, Kamali MR (2006) Geochemistry and origin of natural gas in the world's largest non-associated gas field. J Pet Sci Eng 50:163–175

Balkin SD, Ord K (2000) Automatic neural network modeling for univariate time series. Int J Forecast 16(4):509–515

Bishop CM (2006) Pattern recognition and machine learning. Springer Science Business Media, New York

Bolandi V, Kadkhodaie-Ilkhchi A, Alizadeh A, Tahmorasi J, Farzi R (2015) Source rock characterization of the Albian Kazhdumi formation by integrating well logs and geochemical data in the Azadegan oilfield, Abadan plain, SW Iran. J Pet Sci Eng 133:167–176

Brazdil P, Giraud-Carrier C, Soares C, Vilalta R (2009) Metalearning: application to data mining. Springer, Berlin

Breiman L (1996) Bagging predictors. Mach Learn 26(2):123-140

Bromba M, Ziegler H (1979) Efficient computation of polynomial smoothing digital filters. Anal Chem 51:1760

Brown G (2010) Encyclopedia of machine learning. In: Sammut C, Webb GI (eds) Springer Press, New York

Chiu SL (1994) Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 2:567–580

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proc. 13th Intl. Conf. Mach. Learn., pp 148–156

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. Ann Stat 28(2):337–407

Galimove EM, Rabbani AR (2001) Geochemical characteristics and origin of natural gas in southern Iran. Geochem Int 39(8):780–792

Ghazban F (2007) Petroleum geology of the Persian Gulf. Join publication, Tehran University Press and National Iranian Oil Company, Tehran

Hastie T, Tibshirani R, Friedman J (2008) The Elements of statistical learning, second edition. Springer, Berlin

Hunt JM (1996) Petroleum geochemistry and geology, 2nd edn. Freeman and Company, New York

Insalaco E, Virgone A, Courme B, Gaillot J, Kamali M, Moallemi A, Lotfpour M, Monibi S (2006) Upper Dalan Member and Kangan Formation between the Zagros Mountains and offshore Fars, Iran: depositional system, biostratigraphy and stratigraphic architecture. GeoArabia 11:75–176

Jones RW (1987) Organic Facies. In: Brooks J, Welte D (eds) Advances in petroleum geochemistry, vol 2. Academic Press, London, pp 1–90

Kadkhodaie-Ilkhchi A, Rezaee MR, Moallemi S (2006) A fuzzy logic approach for the estimation of permeability and rock types from conventional well log data: an example from the Kangan reservoir in Iran offshore gas field, Iran. J Geophys Eng 3:356–369

Kadkhodaie-Ilkhchi A, Rahimpour-Bonab H, Rezaee MR (2008) A committee machine with intelligent systems for estimation of total organic carbon content from petrophysical data: an example from the Kangan and Dalan reservoirs in south pars gas field, Iran. Comput Geosci 35:459–474

Kadkhodaie-Ilkhchi A, Monteiro ST, Ramos F, Hatherly P (2010) Rock recognition from MWD data: a comparative study of boosting, neural networks, and fuzzy logic. IEEE Geosci Remote Sens Lett 7(4):680–684

Khan AU, Bandopadhyaya TK, Sharma S (2008) Genetic algorithm based back propagation neural network performs better than backpropagation neural network in stock rates prediction. IJCSNS Int J Comput Sci Netw Secur 8(7):162–166

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley. pp 281–297

Monteiro ST, Murphy RJ, Ramos F, Nieto J (2009) Applying boosting for hyperspectral classification of ore-bearing rocks. IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp 1–6

Nikravesh M, Aminzadeh F, Zadeh LA (2003) Soft computing and intelligent data analysis in oil exploration. Elsevier Science, New York

Orfanidis S (2010) Introduction of signal processing. Prentice Hall, New York

Rokach L (2010) Pattern classification using ensemble methods. World Scientific Publishing, Singapore

Saggaf MM, Nebrija EdL (2003) Estimation of missing logs by regularized neural networks. AAPG Bull 87(8):1377–1389

Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 36(8):1627–1639

Schapire R (1999) A brief introduction to boosting. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp 1–6

Sfidari E, Kadkhodaie-Ilkhchi A, Najjari S (2012) Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems. J Pet Sci Eng 86–87:190–205

Sumathi S, Surekha P (2010) Computational intelligence paradigms: Theory and applications using MATLAB®. Taylor and Francis Group

Tan PN, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison Wesley, pp 769

Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353

Ziegler H (1981) Properties of digital smoothing polynomial (DISPO) filters. Appl Spectrosc 35:88–92