APPLIED GEOPHYSICS, Vol.12, No.2 (June 2015), P. 263-272, 9 Figures.
DOI:10.1007/s11770-015-0490-4

Data mining and well logging interpretation: application to a conglomerate reservoir*

Shi Ning[1,2], Li Hong-Qi♦[1,2], and Luo Wei-Ping[1,2]

Abstract: Data mining is the process of extracting implicit but potentially useful information from incomplete, noisy, and fuzzy data. Data mining offers excellent nonlinear modeling and self-organized learning, and it can play a vital role in the interpretation of well logging data of complex reservoirs. We used data mining to identify the lithologies in a complex reservoir. The reservoir lithologies served as the classification task target and were identified using feature extraction, feature selection, and modeling of data streams. We used independent component analysis to extract information from well curves. We then used the branch-and-bound algorithm to look for the optimal feature subsets and eliminate redundant information. Finally, we used the C5.0 decision-tree algorithm to set up disaggregated models of the well logging curves. The modeling and actual logging data were in good agreement, showing the usefulness of data mining methods in complex reservoirs.
Keywords: Data mining, well logging interpretation, independent component analysis, branch-and-bound algorithm, C5.0 decision tree

Introduction

Since first proposed by Fayyad (1996), data mining has become the main tool to retrieve useful information from large datasets. Data mining is the process of nontrivial extraction of hidden and potentially useful information from large amounts of incomplete, noisy, and fuzzy data. At present, data mining is widely used in many fields such as banking, telecommunications, medicine, and sports. The oil exploration of complex and nonconventional reservoirs requires the use of rigorous interpretation techniques of well logging data to complement conventional interpretation methods.

Aminzadel (2005) pointed out that data mining is critical in transforming oil exploration data into useful knowledge. For instance, Al-Bazzaz (2007) used a neural network approach to model the permeability of the Mauddud-Burgan carbonate reservoir in Kuwait. Guo (2008) used the Relief-F algorithm to select well curves that were sensitive to the flooded layer in flooding level division and then built a disaggregated model for flooded layers on the basis of support vector machines. The model could find the level of the flooded layer; however, Relief-F could not satisfy the interpretation requirements for well logging data. Thus, Relief-F is only used for ranking the well curves according to their sensitivities from high to low instead of providing feature subsets.

Manuscript received by the Editor March 1, 2014 ; revised manuscript received April 27, 2015.
*The research is sponsored by the National Science and Technology Major Project (No.2011ZX05023-005-006)
1. Geophysics and Information Engineering College, China University of Petroleum, Beijing 102249 China.
2.Beijing Key Laboratory of Petroleum Data Mining, China University of Petroleum, Beijing 102249 China.
♦Corresponding author: Li Hong-Qi (hq.li@cup.edu.cn)

The well curves have many differences; thus, the most sensitive three-dimensional curves may not constitute the optimal three-dimensional feature subset. Moreover, the dimensionality of the global optimal subset cannot be confirmed on the basis of the ranking of Relief-F.

In search for low-resistivity gas reservoirs in Junggar Basin, Li (2010) improved the accuracy of gas-reservoir predictions by first selecting the sensitive attributes with a genetic algorithm and then classifying them with the C5.0 algorithm. However, actual well logging data are strongly correlated and do not satisfy the conditional independence assumptions of the C5.0 decision-tree algorithm. If this inconsistency between well logging and data mining algorithms is resolved, the data interpretation will improve.

To identify sedimentary microfacies in Jiyuan, Liu (2011) extracted information from well curves by independent component analysis and then used support vector machines in the classification. The extracted attributes better reflected the sedimentary layers than the original well curves. Nevertheless, not every curve of the independent components reflected the sedimentary microfacies. If feature selection is performed using independent components that preserve the attributes related to the classification, the accuracy of the models built with these attributes will be superior to the accuracy of the models based on all other attributes. Li (2013) puts forward a self-organizing feature map neural network based on the task-driven data mining. With the attempt to solve the problem of complex reservoir identification, the decision tree and support vector machine are used to build the fluid predictive model. Meanwhile, the optimization algorithms inclusive of genetic, grid and quadratic are adopted to optimize the parameters of C-SVC and υ-SVC, such as C, υ and γ, so as to improve the classification performance and generalization ability of the predictive model of support vector machine. Shi (2013) identified the lithology of reservoir containing anhydrite by combining independent component analysis and Naive Bayes algorithm. The independent component analysis can make the well-logging curves meet the independence assumption condition of Naive Bayes algorithm, improving the accuracy of lithology identification. Acar (2014) implement data mining process to the well log data produced by natural gas wells of Degirmenkoy Gas Field. The main goal is to select the appropriate algorithm from the set of NNGE, PA and PART algorithms to find gaseous zones in sandstones.

In conclusion, despite the success of data mining in well logging interpretation, the method is immature because it cannot extract or select well logging data. Furthermore, data mining cannot handle the disagreement between the assumptions in the data mining algorithms and the actual well logging data. This study aims to identify conglomerate reservoirs using rigorous mining of data streams taken from logging data. In this study, we use independent component analysis to extract information from well curves, the branch-and-bound algorithm (BAB) for feature selection, and the C5.0 decision-tree algorithm to build disaggregated models.

Data mining of well logs

Following the definition of Fayyad (1996), we establish rules for mining data streams and interpreting well log data. In general, the mining of data streams is based on understanding the specific business use and the nature of the data, and it is followed by data processing, modeling, evaluation, and application.

Well curves reflect the geophysical features of the rocks, and data mining extracts and then selects the geophysical features. Well curves are strongly correlated, which contradicts the conditional independence assumptions of several data mining algorithms, e.g., Bayesian discrimination and C5.0 decision tree. Thus, feature extraction is used to improve the independency of the curves. The dimensionality of the well curves is low, which helps to find the global optimal subset. Compared with mining of large datasets, well log interpretation mainly relies on a small number of curves. For instance, typically, nine conventional curves are used. Feature selection is realized in two steps. In the first step, we find the optimal feature subset of each dimension by using BAB. In the second step, we find the global optimal feature subset on the basis of the accuracy of the disaggregated model. Because the dimensionality of well log data is low, the efficiency of BAB is high. Figure 1 shows the basics of data mining, which constitutes of understanding the business use, data preprocessing, feature extraction, feature selection, modeling, and model evaluation and application.

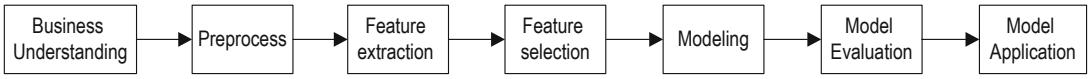

Fig.1 Data mining stream.

The business-understanding step converts the problem of well log interpretation to the task of data mining. For instance, the identification of lithologies is a classification task that is processed with a classification algorithm. Data preprocessing mainly refers to the standardization of the well logging data, the restoration of the core data, elimination of outliers, and others. In feature extraction, new attributes are obtained by combining linearly or nonlinearly the original well curves and the new attributes reflect the original data better than the well curves. The step of feature selection involves the selection of feature subsets using all the attributes and finds the most sensitive attributes while removing the irrelevant ones. In the modeling step, we build a model or models of well logging data by selecting suitable data mining algorithms in line with the requirements of the interpretation. In the model evaluation step, multiple models are evaluated on the basis of the prediction accuracy and the model with the best performance is selected. In the model application, we apply the finally selected model to the actual well log data and examine the model efficiency. The three most important steps in the data mining process are feature extraction, feature selection, and modeling. Feature extraction is performed before feature selection because not every extracted feature is significant to modeling, whereas feature selection finds the feature subsets that are most useful to modeling. Independent component analysis can be used as the feature extraction algorithm to remove the correlation of the well curves and force the data to satisfy the assumptions of classification algorithms, such as C5.0. BAB is used to find the optimal feature subset and the C5.0 decision-tree algorithm is used in the modeling step.

Based on this, we build a data mining process for a conglomerate reservoir using independent component analysis for feature extraction, BAB for feature selection, and the C5.0 decision-tree algorithm for modeling the reservoir lithology.

Feature extraction and independent component analysis

Several classification data mining algorithms, such as Bayesian Network and C5.0 decision tree, require uncorrelated input attributes; thus, the strongly correlated well curves does not satisfy this condition. Therefore, direct modeling of well curves may affect the accuracy of the model. Principal component analysis only extracts the irrelevant signals, whereas independent component analysis extracts the mutually independent signals, and this satisfies the conditional independence assumption required by the classification algorithm.

Independent component analysis decomposes the independent components in the observation signals obtained from multichannel measurements. We assume that $\mathbf{S(t)} = [S_1(t), \ldots, S_N(t)]^T$ is an N-dimensional vector consisting of the original signal source and mutually independent components $S_i(t)$, where ($i = 1, ..., N$) . $\mathbf{S(t)}$ is transformed to the M-dimensional observation vector $\mathbf{X(t)} = [X_1(t), \ldots, X_M(t)]^T$ through the linear combination of the hybrid system $\mathbf{A}$ and noise $\mathbf{N(t)}$. Its expression is

$$\mathbf{X(t)} = \mathbf{A} \cdot \mathbf{S(t)} + \mathbf{N(t)}. \quad (1)$$

For very low noise, $\mathbf{N(t)}$ is zero and thus

$$\mathbf{X(t)} = \mathbf{A} \cdot \mathbf{S(t)}. \quad (2)$$

To set the separation matrix $\mathbf{W} = (w_{ij})$, the observation signal $\mathbf{X(t)}$ is multiplied with the separation matrix $\mathbf{W}$ to obtain vector $\mathbf{Y(t)} = [Y_1(t), \ldots, Y_N(t)]^T$. It also can be expressed by

$$\mathbf{Y(t)} = \mathbf{W} \cdot \mathbf{X(t)} = \mathbf{WA} \cdot \mathbf{S(t)}. \quad (3)$$

Obviously, when $\mathbf{WA} = \mathbf{I}$ ($\mathbf{I}$ is an $N \times N$ -dimensional unit matrix), $\mathbf{Y(t)} = \mathbf{S(t)}$. Vector $\mathbf{Y(t)}$ is the source signal $\mathbf{S(t)}$, i.e., as long as the separation matrix $\mathbf{W}$ is obtained, mutually independent source signals $\mathbf{S(t)}$ can be extracted from the observation signal $\mathbf{X(t)}$.

In independent component analysis, the observation signal is obtained by combining several independent signals linearly; therefore, the decomposed components may have physical significance. For instance, independent component analysis is carried out using porosity and acoustic travel time *AC*, and porosity and compensated neutron logging *CNL*, i.e., $\mathbf{X(t)} = [AC(t), CNL(t)]^T$. Hence, the two independent components *IA* and *IB* are obtained, i.e., $\mathbf{Y(t)} = [IA(t), IB(t)]^T$.

As shown in Figure 2, the correlation ($R^2 = 0.808$) between the independent component *IA* and porosity is higher than between porosity and *AC* ($R^2 = 0.659$) and between porosity and CNL ($R^2 = 0.411$), which suggests that *IA* better reflects porosity. In contrast, in Figure 2d, the correlation between the independent component *IB* and porosity is very low. Therefore, after extracting the independent components from the well curves, feature selection is performed according to the interpretation objectives.

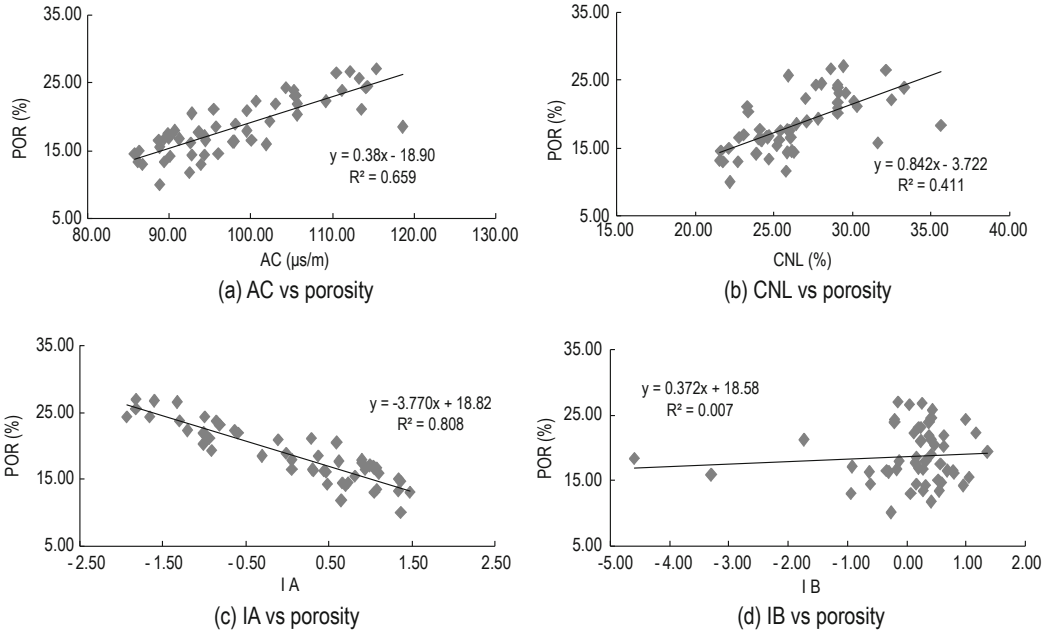

Fig.2 Correlation between independent components and porosity.

Feature selection and the branch and bound algorithm

BAB searches the entire solution space to find the optimal solution. During execution, the algorithm constructs a top-down search tree that constantly divides the feasible solution space into increasingly smaller subsets, which is referred to as branching, and calculates the separation criteria and boundary values for each subset. After each branching step, the magnitude of the separation criteria and boundary values are compared and only subsets with criteria larger than the boundary values are searched. Branching and bounding operate continuously until the maximum feasible solution is obtained.

The separation criteria are functions of the dispersion within and among classifications. Obviously, for distinguishing classifications, the smaller the dispersion of data with the same classification is, the larger the dispersion of data with different classifications is. This allows to better distinguish between different classifications. We assume c sample classifications in N-dimensional space. $\mathbf{x}_k^{(i)}$ is the vector of the ith classification, n_i is the sample size of the ith classification, and p_i is the prior probability of the ith classification. Then, $\mathbf{m}_i$ is the mean vector of the ith sample and $\mathbf{m}$ is the total mean vector of all sample classifications (Held and Karp, 1970):

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}\mathbf{x}_k^{(i)}, \tag{4}$$

$$\mathbf{m} = \sum_{i=1}^{c} p_i \cdot \mathbf{m}_i. \tag{5}$$

$\mathbf{S_b}$ is the dispersion matrix among classifications:

$$\mathbf{S_b} = \sum_{i=1}^{c} p_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^{\mathrm{T}}. \tag{6}$$

$\mathbf{S_w}$ is the dispersion matrix within classifications:

$$\mathbf{S_w} = \sum_{i=1}^{c} p_i \frac{1}{n_i}\sum_{k=1}^{n_i}(\mathbf{x}_k^{(l)} - \mathbf{m}_i)(\mathbf{x}_k^{(l)} - \mathbf{m}_i)^{\mathrm{T}}. \tag{7}$$

$J(x)$ is the separability criterion:

$$J(x) = \frac{\mathrm{tr}(\mathbf{S_b})}{\mathrm{tr}(\mathbf{S_w})}. \tag{8}$$

The search tree starts from the root node and the child node deletes an attribute at every downward search level, until it reaches the leaf node. In the case of an N-dimensional group of original features, we select d attributes to form a feature subset and then the level of the search tree becomes $N-d$. The zeroth level of the search tree is the root node, including all the N attributes. The $(N-d)$th level of the tree is the leaf node and each leaf node includes d attributes. Using Figure 3 as an example, we demonstrate the selection of two attributes from

the six-dimensional feature group $X_0 = \{x_1, x_2, \ldots, x_6\}$. The number at each node represents the rejected attributes. For instance, node A rejects the x_3 attribute. The search tree has four levels altogether, where the node of the fourth level is the leaf node and i denotes the level of the search tree. The search process starts at the root node from left to right. The so-called ith level of the node under discussion refers to the rightmost node at the ith level that was not branched, and the search tree branches from this node. Set $\overline{X_i}$ represents all the attributes of the current node at the ith level. $\psi_i = \{x_1, x_2, \ldots, x_{qi}, x_{ri}\}$ is the alternative feature set that represents the rejected attributes at the child nodes of the current node in Set $\overline{X_i}$, and its feature number is r_i. The number of the current nodes is $q_i = r_i - (N-d-i-1)$. $q_i = 0$ indicates that there is no child node at this node, i.e., this is a leaf node. Thus, the algorithm needs to trace backward. For $q_i > 0$, the current node is not a leaf node and the algorithm continues to search downward. At the current node, according to the alternative feature set ψ_i, the algorithm calculates the $J(\overline{X_i} - x_j)$ separation criterion of r_i rejecting schemes, where $x_j \in \psi_i$. We then rank the criteria in line with the magnitude of J and select q_i minimum J values. The corresponding attributes x_j are successively rejected at each child node from left to right, i.e., among the child nodes of the current node, the criterion for the child node on the left is always less than that of the child node on the right. Among the child nodes, the rejections make up set $Q_i = \{x_1, x_2, \ldots, x_{qi}\}$ for $\overline{X_i} \geq \psi_i \geq Q_i$.

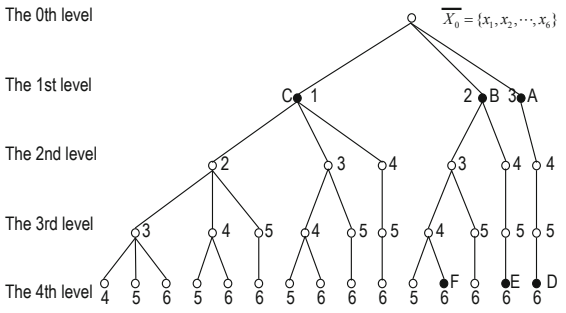

Fig.3 Tree structure of the branch-and-bound algorithm.

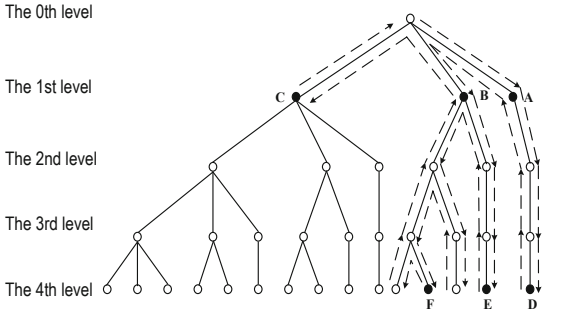

Fig.4 Path search of the branch-and-bound algorithm.

Bounding mainly uses the monotonicity of the separation criteria. If the separation criteria of a feature set are smaller than the boundary values, then the criteria for any subset will also be smaller than the boundary values. After each branching, we compare the magnitude of the separation criteria and boundary values of the current node. We do not branch subsets with separability criteria smaller than their boundary values and continue searching subsets with separation criteria larger than their boundary values. We consider the boundary value B_d, which is the maximum among all the J values of the known leaf nodes. When the search tree moves downward, it calculates the J value of the new node and compares it with the boundary value B_d to establish the search direction. If $J < B_d$, it moves back to the rightmost unsearched node; otherwise, if $J > B_d$, it continues to search downward. The entire bounding process starts from the rightmost branch and ends at the leftmost branch of the search tree. When it ends, the leaf node that B_d corresponds to is the optimal feature subset. Using Figure 4 as an example, we assume that the order of the separability criteria of the black nodes is $J_A > J_B > J_F > J_C > J_D > J_E$. Then, the search moves along the dotted line in Figure 4 and B_d corresponds to the leaf node F. Consequently, the separation criterion of node F is maximal and the preserved attribute is the optimum for classification.

After establishing the dimensionality of the feature subsets, BAB finds the optimal subset and compares the classification of the optimal feature subsets of different dimensions to find the global optimal solution. The nine conventional well curves constitute 512 feature subsets. If BAB is adopted, the optimal solution is found by comparing only the nine feature subsets, which is highly efficient. Moreover, the algorithm avoids inappropriate selections of the dimension of the feature subset.

Building disaggregated models with the C5.0 decision tree

The C5.0 decision-tree algorithm is commonly used to construct a tree-shaped disaggregated model to reflect the relation between classifications and data attributes. The decision tree compares the magnitude of the attributes at the internal nodes using a top-down, divide-and-conquer recursive approach, evaluates the

branching scheme for the next step, and finally obtains the classification at the leaf node. In the decision-tree model, every path from the root node to the leaf node corresponds to one classification rule, and the whole tree corresponds to a group of classification rules.

The C5.0 algorithm uses the information gain ratio as partition metric and selects the maximum attribute of the information gain ratio as the division attribute of the current node. We assume that set S of sample size n is partitioned into c different classifications C_i(i=1, 2,…, c). Each classification C_i contains n_i samples and the information entropy of S is

$$E(S) = -\sum_{i=1}^{c} p_i \cdot \log_2(p_i), \tag{9}$$

where p_i is the prior probability of the ith classification, i.e., $p_i = n_i/n$. We assume that the A attributes are all X_A = $\{A_1, A_2, \ldots, A_v\}$ and divide set S into v subsets $\{S_1, S_2, \ldots, S_v\}$ on the basis of the A values (Quinlan, 1986). The A values of the samples in each subset are all the same. For instance, the A values of the samples in subset S_v are A_v and the conditional entropy of the A attributes is

$$E(S, A) = \sum_{i \in X_A} \frac{|S_i|}{|S|} E(S_i), \tag{10}$$

where $E(S_i)$ is the information entropy of subset S_i.

Information gain is the difference between information entropy and conditional entropy, and the information grain of A attribute is

$$Gain(A) = E(S) - E(S, A). \tag{11}$$

The information gain ratio is the improved form of the information gain, and it considers not only the magnitude of the information gain but also the cost to obtain the information gain. The information gain ratio of the A attribute is

$$GainR(A) = \frac{Gain(A)}{-\sum_{j=1}^{v} \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|}}. \tag{12}$$

The C5.0 algorithm adopts conservative error pruning and uses the error rate of the leaf node on the training set to estimate the error rate of the upper limit of the unknown data on the basis of specific confidence level. The unknown data are deemed to follow the binomial distribution and the errors are eliminated accordingly.

Application

Geological background

The conglomerate reservoir in Junggar Basin is an alluvial fan deposit that consists of conglomerate, glutenite, sandstone, argillaceous siltstone, and mudstone. The lithologies of terminal and middle fan subfacies are coarse-grained conglomerate and glutenite, whereas the fan edge subfacies are fine-grained sandstone, argillaceous siltstone, and mudstone. The well curves are the spontaneous potential *SP*, natural gamma rays *GR*, acoustic travel time *AC*, density *DEN*, compensated neutron logging *CNL*, true formation resistivity RT, invaded zone resistivity *RI*, and flushed zone formation resistivity *RXO*. Table 1 lists the well logging data for each lithology. The data overlap probably because of the poor sorting of the rocks. Moreover, the well curves are fuzzy and the lithologies are difficult to distinguish. Thus, we adopt the data mining method and the three steps of feature extraction, feature selection, and modeling. The adopted algorithms are independent

Table 1 Well logging data for different lithologies

Lithology		*GR* API	*SP* mV	*AC* μs/m	*DEN* g·cm^{-3}	*CNL* %	*RT* Ω·m	*RI* Ω·m	*ROX* Ω·m
Conglomerate	Max	64.12	−7.80	121.75	2.42	37.03	72.32	64.27	86.96
	Min	47.20	−32.26	90.04	2.29	23.92	7.45	7.13	6.52
Mudstone	Max	62.71	−5.61	121.73	2.50	42.17	71.69	59.32	51.66
	Min	38.50	−24.53	90.85	2.30	24.48	7.15	6.79	8.39
Argillaceous siltstone	Max	55.70	−6.70	116.91	2.47	39.73	57.80	50.29	54.96
	Min	46.20	−31.95	99.58	2.30	27.25	7.77	7.09	7.73
Glutenite	Max	55.60	−7.65	128.75	2.45	43.88	93.98	85.76	99.65
	Min	21.80	−33.44	88.77	2.17	22.19	5.98	4.31	4.13
Sandstone	Max	62.40	−8.56	115.35	2.36	32.14	74.97	61.17	109.31
	Min	42.80	−26.65	104.46	2.28	25.90	19.33	16.96	29.88

component analysis, BAB, and the C5.0 decision tree. Preprocessing serves to standardize the curves and restore the core data before feature extraction.

Feature extraction

The objective of feature extraction is to extract new features from the original data. The curves are strongly correlated and do not satisfy the independence assumptions of the C5.0 decision-tree algorithm (see Table 2). The independent component analysis of the well curves serves to extract new attributes.

Table 2 Correlation coefficient matrix of the logging data

	GR	*SP*	*AC*	*DEN*	*CNL*	*RT*	*RI*	*RXO*
GR	1							
SP	−0.04	1						
AC	−0.45	0.18	1					
DEN	0.35	0.39	−0.57	1				
CNL	−0.36	0.44	0.74	−0.12	1			
RT	0.10	−0.76	−0.45	−0.20	−0.69	1		
RI	0.14	−0.71	−0.57	−0.10	−0.75	0.96	1	
RXO	0.02	−0.53	−0.34	−0.10	−0.65	0.75	0.77	1

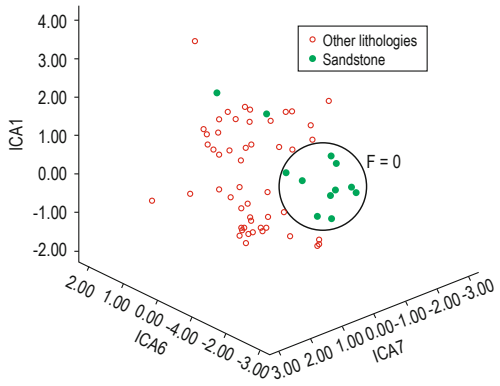

Fig.5 Independent component curves to distinguish sandstone and other rock.

Figure 5 shows that $IC1$, $IC6$, and $IC7$ can better distinguish sandstone (green solid dots) from other lithologies. Most sandstone values are characterized by $F \le 0$, where

$$F = \sqrt{(IC1-0.05)^2 + (IC6+1.5)^2 + (IC7-0.46)^2} - 1.54.$$

Figure 6 shows that $IC1$, $IC2$, and $IC6$ better distinguish coarse-grained rocks (red solid circles) than fine-grained rocks (green open circles). For $G \le 0$, the rocks are coarse, whereas the rocks are fine for $G > 0$, where

$$G = 0.94 \cdot IC2 - 0.75 \cdot IC1 - 1.87 \cdot IC6.$$

From Figures 5 and 6, we see that the independent components extracted from the well curves not only satisfy the conditional independence assumptions of the classification algorithm but also can be used to classify the lithology. Next, we perform feature selection on the basis of the eight curves of independent components to find the most suitable feature subset for identifying the different lithologies.

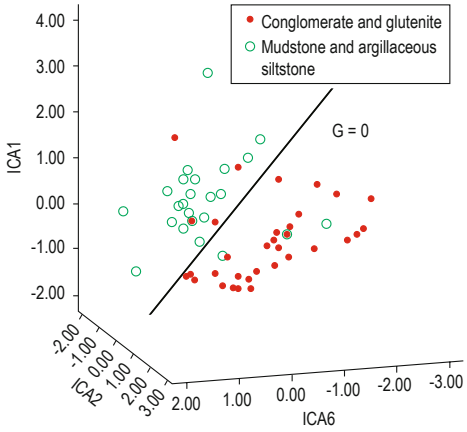

Fig.6 Independent component curves to distinguish lithologies on the basis of grain size.

Feature selection

In the modeling step, the goal of BAB is to find the optimal feature subsets at different dimensions according to the dimensions of the feature subsets. For, the global optimal solution is surely among them. The independent component input is an eight-dimensional variable and the dimensions of the feature subsets range from one to seven; therefore, as long as the optimal feature subset of each dimension is obtained, the overall optimal feature subset can be found by combining the eight-dimensional universal set. Using the five-dimensional feature selection as an example, we classify the five lithologies, where $c = 5$; $N = 8$, and $d = 5$. We calculate the separation criterion of the sample using Equation 8 and finally obtain the five-dimensional optimal feature subsets $IC1$, $IC2$, $IC6$, $IC7$ and $IC8$. Table 3 lists the optimal feature subsets for the different dimensions. Disaggregated models are built for each feature subset in Table 3 and the global optimal feature combination is found using the accuracy of the models.

Model building

We use the C5.0 decision-tree algorithm to identify the different lithologies, using the original core data

as input, five classification samples, seven optimal subsets, and the universal set of the ICA curve as input attributes. During each modeling step (eight total), we only adjust the input feature subsets and construct eight ICA disaggregated models. To compare the ability of the curves of the independent components to distinguish lithologies and that of the well curves, disaggregated models of the well curves at different dimensions are built using the branch-and-bound and the C5.0 decision-tree algorithms.

Table 3 Feature subsets of the ICA curves

Feature subset	1-Dsubset	2-D subset	3-D subset	4-D subset	5-D subset	6-D subset	7-Dsubset
	*IC*6	*IC*6	*IC*1	*IC*1	*IC*1	*IC*1	*IC*1
		*IC*7	*IC*6	*IC*6	*IC*2	*IC*2	*IC*2
			*IC*7	*IC*7	*IC*6	*IC*4	*IC*4
				*IC*8	*IC*7	*IC*6	*IC*6
					*IC*8	*IC*7	*IC*7
						*IC*8	*IC*5
							*IC*8

Table 4 Comparison of the model accuracy for different feature subsets

Feature subset	1D subset %	2D subset %	3D subset %	4D subset %	5D subset %	6D subset %	7D subset %	universal set %	Mean %
ICA curves	56.00	68.00	80.00	82.00	90.00	84.00	82.00	82.00	77.50
Well curves	52.00	66.00	64.00	82.00	80.00	80.00	82.00	80.00	73.00

Table 4 lists the accuracy of the ICA model and the model of the well curves. Two points are of interest. First, the average accuracy of the ICA model is higher than that of the well curves, as it complies with the assumed conditions of the C5.0 decision-tree algorithm. Second, for the disaggregated models, increasing the number of input features does not improve the accuracy. For both well and ICA curves, the model with the highest accuracy is not the universal set; moreover, BAB finds the feature subset that best fits the classification. Among all the models, the model built with the five-dimensional feature subset (*IC*1, *IC*2, *IC*6, *IC*7, and *IC*8) of the ICA curve has the highest accuracy (90.00%). Figure 7 shows the five-dimensional decision-tree model of the ICA curve, where the elliptic nodes are internal nodes and marked at the nodes are the tested attributes at each node. Each branch represents a test result. Rectangular nodes are leaf nodes, which represent the classifications evaluated by the model. Each path from the root node to the leaf node corresponds to a classification rule. For instance, the rule at the leftmost of the model is the following. When *IC*1 > 0.17, *IC*6 > 0.55, and *IC*8 > 0.02, the sample is mudstone. In applying the model, each new test starts at the tree root (i.e., *IC*1) and moves down from the top, and its lithology is assessed at the leaf node.

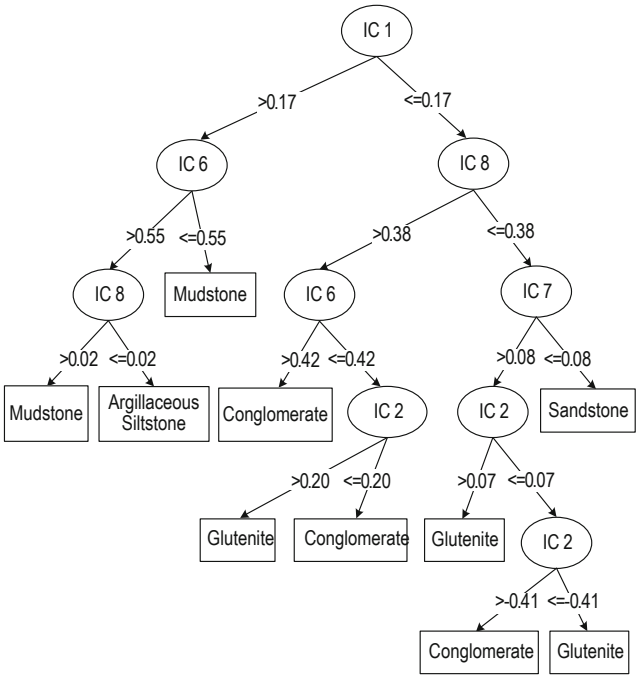

Fig.7 Decision-tree model based on the five-dimensional feature subset of the ICA curve.

Model testing

The lithology-identification ability of the data-mining model is tested using two wells that we did not use to build the model. Figure 8 shows the well A cross section. The three leftmost tracks are well curves, the fifth to seventh tracks are independent component curves, the eighth track is the cross section of the lithology obtained

with the model, and the ninth track represents the actual core data. Clearly, the model-derived lithology is consistent with the actual core data.

Figure 9 shows the cross section of well B. The model-derived lithology is consistent with the original core data.

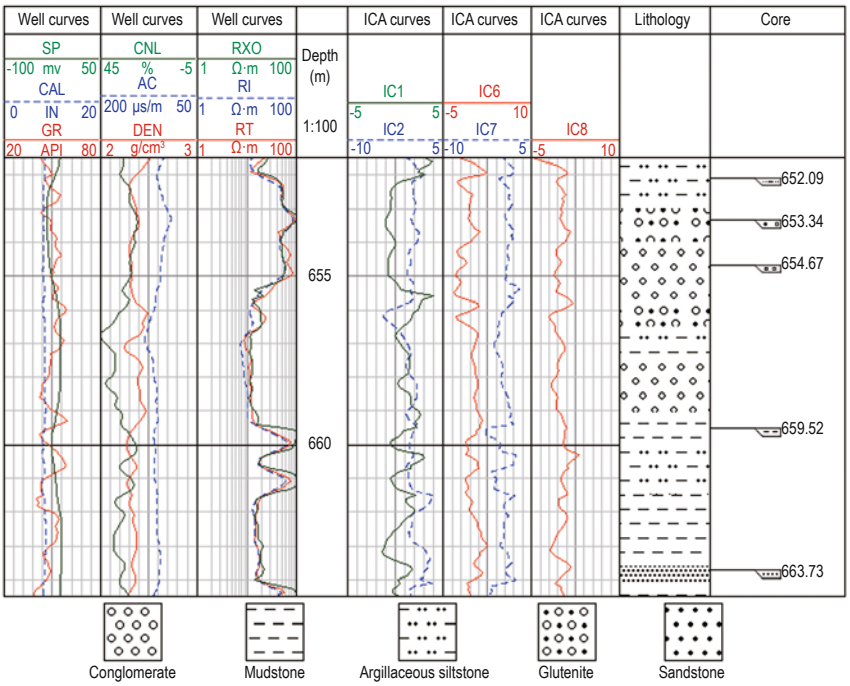

Fig.8 Lithology identification results for well A.

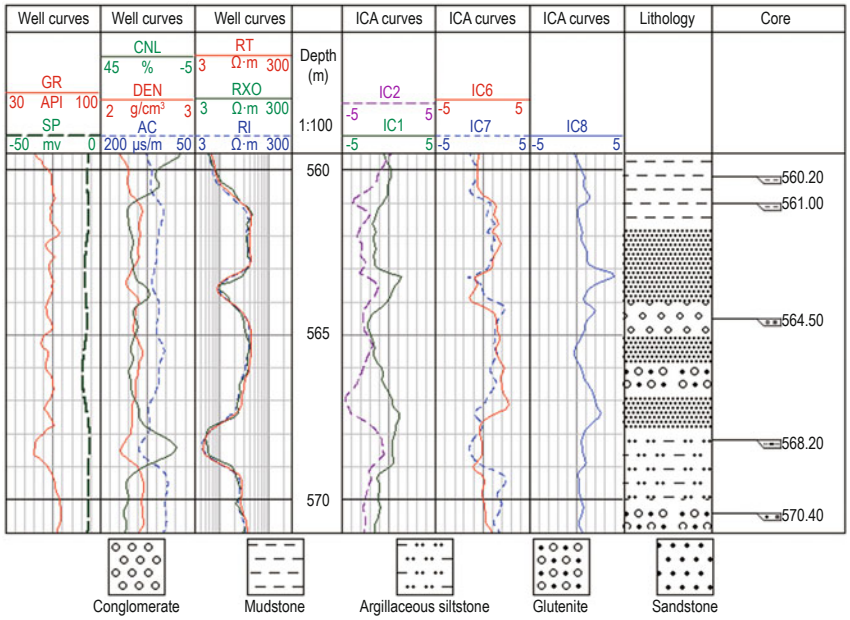

Fig.9 Lithology identification results for well B.

Conclusions

We used the data-mining approach to model the lithology of a complex reservoir with good results. Because of the strong correlation of the well curves, we used independent component analysis for feature extraction, BAB for feature selection, and the C5.0

decision-tree algorithm for modeling.

Acknowledgments

We wish to thank the anonymous reviewers for their constructive comments and suggestions that greatly improved the manuscript.

References

Acar, M. A.,Tolun, M. R., and Elbasi, E., 2014, An application of data mining and knowledge discovery process in the field of natural gas exploration: Proceedings on Application of Information and Communication Technologies (AICT), 8th IEEE International Conference, 1−6.

Al-Bazzaz, W. H., and Al-Mehanna, Y. W., Gupta, A., et al. 2007, Permeability modeling using neural-network approach for complex Mauddud-Burgan carbonate reservoir: SPE Middle East Oil and Gas Show and Conference, 11−14.

Aminzadeh, F., 2005, Applications of AI and soft computing for challenging problems in the oil industry: Journal of Petroleum Science and Engineering, **47**(1−2), 5−14.

Bian, Z. Q., and Zhang, X. G., 2002, Pattern Recognition: Tsinghua University Press, China, 176−210.

Fayyad,U.,and Piatetsky-Shapiro, G., et al. 1996, From data mining to knowledge discovery in databases: AI Magazine, **17**(3), 37−53.

Guo, H. F., Li, H. Q., and Meng, Z. X., 2008, Feature Selection, Genetic Algorithm and Support Vector Machine: Journal of Oil and Gas Technology, **30**(6), 94−99.

Han, J. W., and Kamber, M., 2006, Data mining concepts and techniques: China Machine Press, China, 184–224.

Held, M., and Karp, R. M., 1970, The traveling-salesman problem and minimum spanning trees: Operations Research, **18**(6), 1138−1162.

Jiang, H. L., and Zhao, Q. L., 2009, Machine learning methods: Publishing House of Electronics industry, China, 20−56.

Jutten, C., and Herault, J., 1988, Independent Component Analysis versus PCA. Proceeding of European Signal Processing Conf, 287–314.

Li, H. Q., Li, X. Y., and Tan, F. Q., et al. 2010, Lithology Identification of Conglomerate Reservoir Based on Decision Tree Method: Well Logging Technology, **34**(1), 16–21.

Li, X. Y., 2011, Methods and Applications of Data Mining in the Reservoir Evaluation: PhD Thesis, University of Petroleum, Beijing.

Li, X. Y., and Li, H. Q., 2013, A new method of identification of complex lithologies and reservoirs: task-driven data mining: Journal of Petroleum Science and Engineering, **109**, 241–249.

Liaqat Ali., and Sandip Bordoloi., et al. 2008,Modeling permeability in tight gas sands using intelligent and innovative data mining techniques. SPE Annual Technical Conference and Exhibition, 21−24.

Liu, J., Li, Z. C., and Wang, Z., 2011, Quantitative identification of microfacies based on ICA,PCA and SVM: Well Logging Technology, **35**(3), 262−265.

Qin, F., Ren, S. L., and Cheng, Z, K., et al. 2007, Bayes classification model based on ICA: Computer Engineering and Design, **28**(20), 4873−4877.

Quinlan, J. R., 1986, Induction of decision trees: Machine learning, **1**(1), 81−106.

Shi, N., and Li, H. Q., and Luo, W. P., 2013, Lithology identification of reservoir containing anhydrite based on independent component analysis and Naive Bayes algorithm: Journal of Xi'an Shiyou University (Natural Science Edition), **28**(5), 39−42.

Yang, F. S., and Hong, B., 2006, Independent component analysis principles and applications: Tsinghua University Press, China, 91−111.

Yong, S. H., and Zhang, C. M., 2006, Well logging data processing and integrated interpretation: China University of Petroleum Press, China, 533−566.

Zhong, Y. H., and Li, R., 2009, Application of Principal Component Analysis and Least Square Support Vector Machine to Lithology Identification: Well Logging Technology, **33**(5), 425−429.

Shi Ning is a PhD student at the China University of Petroleum (Beijing). He received his MS in 2010 from Yangtze University. His main research interests are well logging interpretation and data mining.