Comput Geosci
DOI 10.1007/s10596-016-9601-4

ORIGINAL PAPER

Evaluation of complex petroleum reservoirs based on data mining methods

Fengqi Tan[1,2] · Gang Luo[1,2] · Duojun Wang[1,2] · Yangkang Chen[3]

Received: 20 May 2016 / Accepted: 25 October 2016

Abstract In this study, we introduce the application of data mining to petroleum exploration and development to obtain high-performance predictive models and optimal classifications of geology, reservoirs, reservoir beds, and fluid properties. Data mining is a practical method for finding characteristics of, and inherent laws in massive multi-dimensional data. The data mining method is primarily composed of three loops, which are feature selection, model parameter optimization, and model performance evaluation. The method's key techniques involve applying genetic algorithms to carry out feature selection and parameter optimization and using repeated cross-validation methods to obtain unbiased estimation of generalization accuracy. The optimal model is finally selected from the various algorithms tested. In this paper, the evaluation of water-flooded layers and the classification of conglomerate reservoirs in Karamay oil field are selected as case studies to analyze comprehensively two important functions in data mining, namely predictive modeling and cluster analysis. For the evaluation of water-flooded layers, six feature subset schemes and five distinct types of data mining methods (decision trees, artificial neural networks, support vector machines, Bayesian networks, and ensemble learning) are analyzed and compared. The results clearly demonstrate that decision trees are superior to the other methods in terms of predictive model accuracy and interpretability. Therefore, a decision tree-based model is selected as the final model for identifying water-flooded layers in the conglomerate reservoir. For the reservoir classification, the reservoir classification standards from four types of clustering algorithms, such as those based on division, level, model, and density, are comparatively analyzed. The results clearly indicate that the clustering derived from applying the standard K-means algorithm, which is based on division, provides the best fit to the geological characteristics of the actual reservoir and the greatest accuracy of reservoir classification. Moreover, the internal measurement parameters of this algorithm, such as compactness, efficiency, and resolution, are all better than those of the other three algorithms. Compared with traditional methods from exploration geophysics, the data mining method has obvious advantages in solving problems involving calculation of reservoir parameters and reservoir classification using different specialized field data. Hence, the effective application of data mining methods can provide better services for petroleum exploration and development.

Keywords Data mining · Feature selection · Performance evaluation · Decision tree · Clustering analysis · Conglomerate reservoir

✉ Fengqi Tan
tanfengqi@ucas.ac.cn

[1] College of Earth Science, University of Chinese Academy of Sciences, Beijing 100049, China

[2] Key Laboratory Computational Geodynamics, Chinese Academy of Sciences, Beijing 100049, China

[3] Jackson School of Geosciences, University of Texas at Austin, Austin, TX, 78713-8924, USA

1 Introduction

With the rapid development in the breadth and depth of petroleum exploration and production, traditional methods are encountering two challenges in how to handle the

Published online: 15 November 2016

rapid accumulation of massive quantities of complicated petroleum data. On the one hand, the amount of data related to petroleum exploration and development is increasing rapidly and continuously; if these big data could be used successfully and fully, it would provide additional benefits to petroleum exploration and production. However, difficulties are often encountered because traditional geophysical exploration methods, which are based on rock physics, mathematics, physics, statistics, and petroleum exploration theory, are not capable of utilizing such massive datasets [1, 2]. On the other hand, unconventional and complex oil and gas reservoirs are becoming major exploration targets [3, 4]. For these types of reservoirs, classic techniques such as cross-plots, linear regression, and multivariate discriminant analysis cannot effectively solve the problems of reservoir parameter calculation and fluid feature identification [5]. Therefore, in order to address the challenges faced by petroleum exploration and development, it is necessary to utilize some new technologies from other research fields, such as artificial intelligence, machine learning, and pattern recognition, and to investigate their specific applications. Neural network techniques have been widely used for rock identification, classification of sedimentary facies, permeability prediction, discrimination of oil from gas or water, etc. [6–8]. It has been shown that the nonlinear neural network method can be superior in practice to the linear statistical analysis technique when solving complicated petroleum geology problems that are controlled and affected by multiple factors [9].

Recently, many intelligent modeling methods, such as decision trees (DTs), support vector machines (SVMs), Bayesian networks (BNs), and ensemble learning (ELs), have been gradually applied in different fields [10–12]. The main principle for solving these problems is to identify a relationship between the input and output parameters from the sample datasets by using learning algorithms [13], and the common characteristics of these intelligent modeling methods include features such as a high degree of parameterization, powerful learning ability, and widespread applicability. The prediction accuracy of these models is directly related to the data size and representativeness of samples, feature selection, model parameter settings, and the methods used for performance evaluation. However, two problems are commonly and easily induced when these intelligent modeling methods are not appropriately used. The first problem is called overfitting [14]. Specifically, if the model is overfitted to the training data, although the training accuracy may be very high, the prediction accuracy will be low when the model is applied to other datasets. The second problem is called underfitting [14]. In that case, the model is insufficiently trained and, hence, does not learn the true structure of the sample data. Both of the above problems have significant negative effects on the performance of intelligent methods in practical applications related to oil and gas exploration and development.

Data mining methods have been recognized as one of ten key technologies in dealing with future challenges in petroleum exploration and development [1]. They have been successfully applied to real-time reservoir monitoring [15], best practice identification in petroleum engineering projects [16], reservoir characterization [17], etc. Based on the processing problems associated with massive data and the technological features of data mining methods, a practical approach to data mining that has not been sufficiently utilized in the petroleum industry is proposed in this study. Compared with previous approaches, the proposed approach emphasizes obtaining a highly generalized practical model. Its key technique involves the use of a genetic algorithm to select character subsets and optimize model parameters in the context of performing a comprehensive analysis of various algorithm principles and features, and then determining the final and optimal model from the various prediction models. The data mining approach proposed in the paper can provide improved service and support for the exploration and development of complicated oil and gas reservoirs.

2 Data mining methods and techniques

Data mining is defined as a set of nontrivial processes for obtaining correct, novel, potentially useful, and understandable patterns from massive data [18]. These patterns can be described as functions, rules, trees, networks, etc. The tasks of data mining can be divided into two main categories, description and prediction. The former can summarize the general patterns of potential relationships in the data, and the latter makes predictions by analyzing and making inferences from the current data. According to the types of patterns detected, the functions of data mining can be divided into six categories, which are concept description, correlation analysis, prediction modeling, cluster analysis, anomaly detection, and evolution analysis. Data mining is a process that involves continuous looping and optimization and includes six steps, which are data selection, data preprocessing, data transformation, data mining, pattern evaluation, and knowledge representation [19]. The process is iterative and interactive, as shown in Fig. 1.

An appropriate dataset for data mining includes a set of data objects (also called records, vectors, patterns, samples, data points, etc.) that can be described by using a set of object attributes (also called variables, characters, fields, parameters, dimensions, etc.) that effectively reflect basic features of these objects. The prediction task is to predict

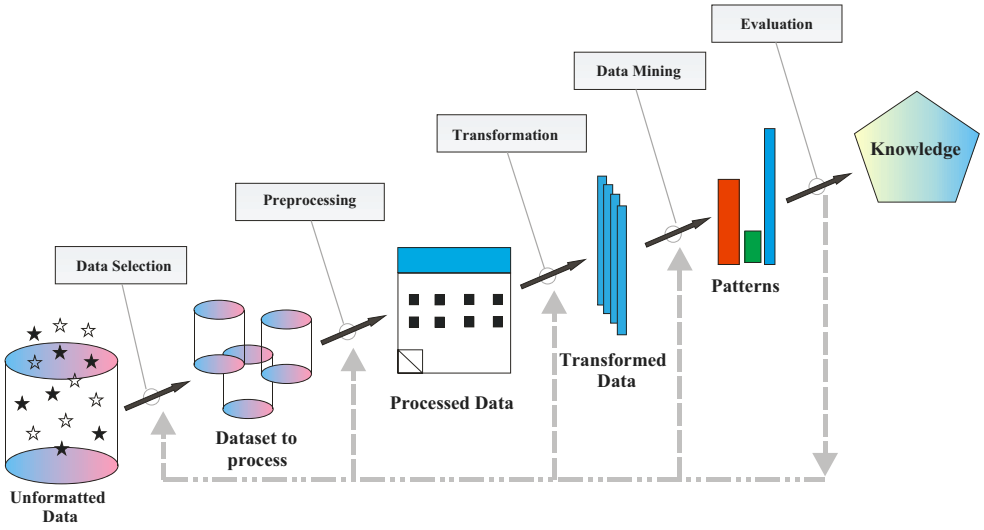


Fig. 1 Flowchart describing the data mining method (Symeonidis and Mitkas 2005)

and estimate the value of a specified attribute by using values of known attributes. The predicted attribute is called the target variable, and the value of the target variable is also called the class label of the target class. Based on the above definition, prediction modeling involves obtaining a target function (f) from the input data with known class labels by way of analysis and learning. Each attribute set (x) is mapped to a predefined class label (y); if y is discrete, this mapping is called classification, and if y is continuous, this mapping is called regression. The generalization ability of a model means the applicability of the model to new sample data. Clustering analysis belongs to the description function of data mining. It makes the original dataset into an object of classification that is divided into different clusters according to different features of the data, and the final purpose is to make all the objects in the same cluster have high similarity, but the objects in different clusters should be substantially different. Generally, the degree of similarity of cluster results is based on the distance, which is defined as the approximate degree between objects in space [14, 18].

The prediction modeling methods used in data mining originally derive from many research fields such as machine learning [20], artificial intelligence [21], pattern recognition [22], and statistics [23] and come out of directed improvements and extensions in these research fields based on the characteristics of data mining. According to the application features of data mining in petroleum exploration and development and the corresponding technical challenges, five approaches to data mining, including decision trees [24, 25], neural networks [22], support vector machines [26, 27], Bayesian networks [28–30], and ensemble learning [31, 32], are studied in this paper. Clustering analysis is an unsupervised learning approach. The presence or absence of supervision is the major difference between clustering analysis and prediction classification [33]. The main technical feature of clustering analysis is that it can automatically classify objects into meaningful clusters by the similarity degree between objects, without knowing the class labels in advance. In this study, four algorithms for cluster analysis, which are based on division, level, model, and density, are investigated and discussed [13, 34]. The mathematical and physical theories underlying these algorithms, as well as the definitions of their parameters and rules for their application, are described in detail in previous studies [14, 34].

3 Data mining approaches in petroleum exploration and development

Data mining is a data-driven approach. This approach does not require geologists and geophysicists to fully understand the mechanisms of rock physics, the rules of variation for reservoir properties, the characteristics of dynamic variation in reservoirs, etc. Instead, one only needs some background knowledge [18]. Using this approach, models can be set up directly from the sample data. After data mining experts analyze and interpret the models, these models can be applied to evaluate complicated reservoirs to obtain satisfactory interpretational success rates in practical applications. To improve the accuracy of prediction models and the reasonableness of clustering analysis results, the problems of overfitting and underfitting must be addressed when data mining methods are used in petroleum exploration and development. Based on the above analysis, the following data mining workflow in petroleum exploration and development is proposed in this study (Fig. 2).

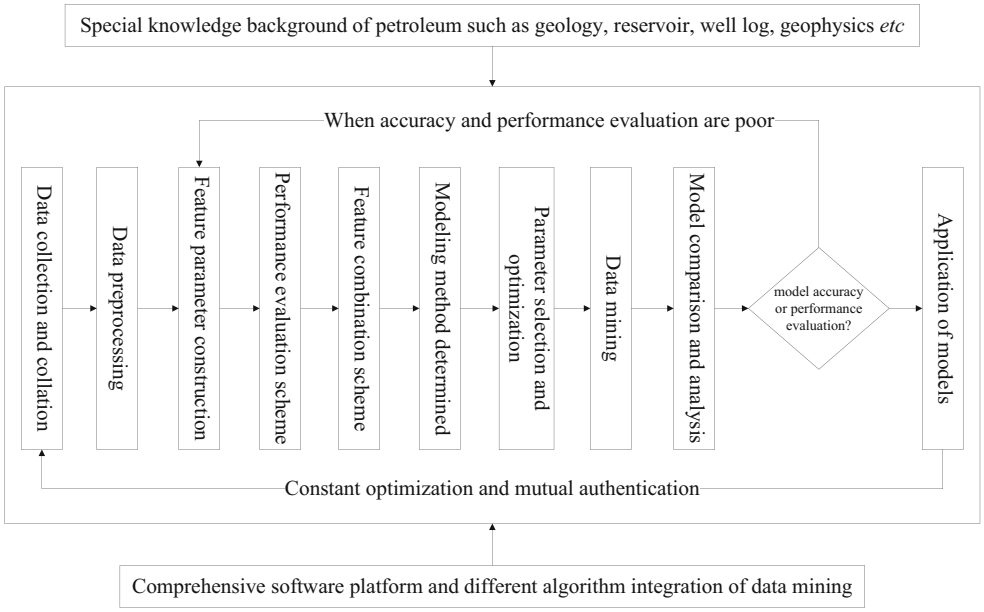


Fig. 2 Technical flowchart describing data mining in petroleum exploration and development

1. Step 1 involves data collection and collation. Based on problems specific to petroleum geology and reservoir development, reliable information and data from various projects must be collected, ensuring data quality as far as possible.
2. Step 2 involves data preprocessing. This step mainly includes four aspects such as the elimination of abnormal data points, supplementing missing values, standardization, and normalization (as needed).
3. Step 3 involves feature parameter construction. It is necessary to construct new parameters for evaluation, depending on the problems to be solved.
4. Step 4 involves the selection of a performance evaluation scheme.
5. Step 5 involves the selection of a feature combination scheme.
6. Step 6 involves the selection of a modeling method. The modeling method combination is chosen based on the difficulties that need to be solved in the petroleum exploration and development problem of interest and the features of the selected algorithms.
7. Step 7 involves performing data mining on the datasets.
8. Step 8 involves model comparison and analysis. Through comparative analysis of the results from various feature subsets and different modeling methods, the optimal model is selected, and this model is analyzed and interpreted based on professional knowledge of petroleum geology, geophysics, reservoir engineering, and other petroleum-related fields.
9. Step 9 involves model application. The selected optimal model is applied to a practical problem. Depending on feedback from the latest results, the model parameters and feature combinations are revised and optimized continuously in order to finally solve practical problems in petroleum exploration and development.

A model's generalization ability is controlled by many factors such as sample representativeness, algorithm selection, and model complexity [13, 18]. Thus, when applying the aforementioned approaches, five key technical points, described below, must be kept in mind.

1. Generalization ability has generally been considered when designing algorithms in many prediction modeling methods, as reflected by pruning operations in decision trees, learning termination in advance in artificial neural networks (ANNs), and the penalty factor (C) in SVMs. Thus, these operations need to be activated, and the corresponding parameterizations should also be optimized before the application of data mining approaches.
2. Overfitting may be caused when redundant and uncorrelated features exist in the data mining process. Therefore, feature selection is very necessary because it can identify comparison schemes for feature selection by using either manual empirical methods or automatic selection methods.
3. The predictive ability of the selected model needs to be accurately estimated and compared. If the size of the

datasets is sufficiently large that they can be divided into independent training and testing sets, the performance can be evaluated separately in the testing set. If the size of the datasets is limited, in order to avoid underfitting due to insufficient data, the datasets should not be divided into independent training and testing sets. Instead, cross-validation methods could be used on all the available data to evaluate the performance and ability of the prediction model.

4. The model is selected based on the principle of Occam's razor [35–37]. Namely, if there are two models with the same generalization error, the simpler model is preferred over the more complicated model because model complexity has a significant effect on the overfitting problem, and the more complicated model will also need to run a rigorous statistical test.
5. After selecting the optimal model, all the data are used to train and regenerate the model, and the final model can be applied in oil fields.

Both feature selection and model parameterization are optimization problems, and they can be solved by using a genetic algorithm [38]. For feature selection based on filtering methods, the performance evaluation of feature subsets is selected as the fitness function. For model parameter selection, the model performance (which usually refers to the error) is taken as the fitness function. Figure 3 shows the key technical steps of the data mining method in petroleum exploration and development. From the figure, we can see that the algorithm needs to be run on the training datasets four times (Fig. 3), that is, in the feature selection loop, the parameterization loop, the performance evaluation loop, and the final model generation. Thus, the whole process is a computationally intensive task that may take much computation time. The model parameters must be carefully selected and determined in order to obtain results in an acceptable time period.

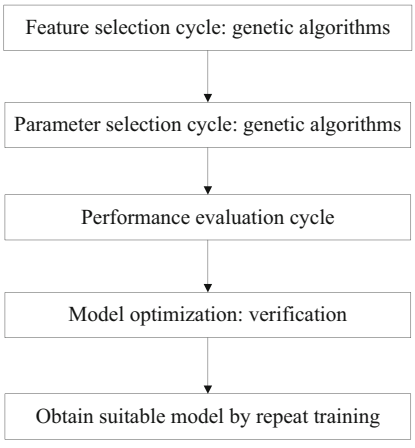


Fig. 3 Key technical steps of the data mining method

4 Case study of prediction modeling

4.1 Geological background in Karamay oil field

The Karamay oil field, which is located at the northwestern margin of Junggar Basin and is developed in the Kexia Group reservoir in Triassic rocks, is a typical conglomerate reservoir. Overall, the sedimentary environment of the conglomerate reservoir in this area belongs to a set of positive cycle piedmont alluvial fan sedimentation which overlaps above the Carboniferous formation. Various subfacies, including top fan, middle fan, and edge fan subfacies, are developed in this oil field [39, 40]. After 50 years of water-injected development, the average overall water content is often above 85 %. Therefore, accurate recognition and quantitative evaluation of water-flooded layers have become difficult and urgent problems in secondary development. However, the special sedimentary environment of conglomerate reservoirs, such as near source, multiple streams, and rapid changes, leads to severe heterogeneities, as well as complicated and variable lithologies (Fig. 4), and pore structure characteristics with complex modes. After water flooding, reservoir heterogeneity and rock matrix can cause a much stronger response than fluid property changes; thus, water flooding conditions cannot be effectively identified from original well logging curves. Various traditional methods of water-flooded layer evaluation cannot obtain good results, and the evaluation accuracy is also too low to meet the requirements of practical applications [41, 42].

4.2 Data collection from water-flooded layers

The sample data for data mining-based evaluation of water-flooded layers comes from the well logging interpretation results of 170 wells that have liquid production profiles in Karamay oil field. These data are sampled in seven sublayers which are labeled S_7^{33}, S_7^{32}, S_7^{31}, S_7^{23}, S_7^{22}, S_7^{21}, and S_7^{1}, and there are 1170 data records in total. Comprehensive analysis of data from the sealing and coring wells, well logging curves, and the conclusions of oil testing and liquid production profiles in the study area has yielded a total of 23 parameters that can indicate water-flooded levels and are selected as data mining objects. There are three water-flooded feature parameters such as the water production rate (f_w), the degree of oil saturation (S_o), and the recovery index (F_{ow}). The recovery index is defined as the ratio of the difference between the original and current oil saturation to

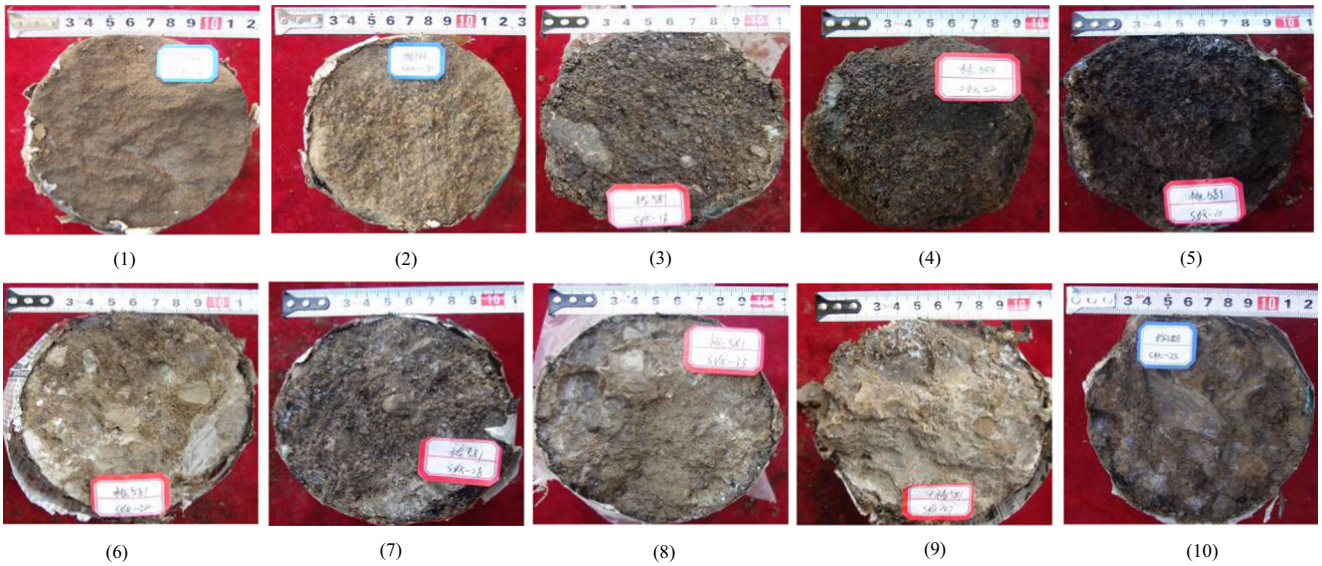


Fig. 4 Lithologies and changes in oil-bearing properties of conglomerate reservoir rocks in Karamay oil field. *1* Gray mudstone, medium. *2* Shale with conglomeratic sandstone, medium, oil patch. *3* Conglomeratic gritstone, medium, oil patch. *4* Conglomeratic gritstone, medium, oil immersed. *5* Conglomeratic gritstone, loose, oil rich. *6* Glutenite, medium, oil patch. *7* Glutenite, loose, oil patch. *8* Conglomerate, tight, oil trace. *9* Conglomerate, medium, oil patch. *10* Conglomerate, medium, oil immersed

the original oil saturation [2], and it is a dynamic parameter describing water flooding in the water injection development process. Higher values of the recovery index reflect stronger flooding, whereas lower values of the recovery index indicate weaker flooding. In addition, the following 20 parameters, including nine derived from traditional well logging curves, lithology, physical properties, fluid properties, constructed parameters from water-flooded layers, formation thickness, and zone information are also selected as variables for prediction modeling of water-flooded layers (Table 1).

4.3 Feature scheme selection

In total, six feature subset schemes are designed to support the process of using data mining to evaluate water flooding (Table 2).

1. The first scheme uses all 23 parameters for model building, without manual or automated selection based on user's experience or intelligent algorithms (No. 1 in Table 2).
2. Three schemes are determined based on user's experience (Nos. 2–4 in Table 2).
3. The fifth scheme is determined based on automated filtering (No. 5 in Table 2). In the filter-based scheme, a genetic algorithm is used to search within the attribute space. Binary chromosome encoding is used to determine the feature subset scheme, and some key parameters are set as follows. In this method, the maximum number of generations is set to 100, the group size is set to 20, the crossing probability is set to 0.5, the variation probability is set to 0.1, the operator is selected by using a roulette wheel method, and then the mutation and uniform crossing operators are adopted.
4. The last scheme is determined (No. 6 in Table 2) by the feature weighting method, also known as the single-factor analysis of variance.

5 Modeling approach and parameterizations

In total, 12 models and algorithms from five methods such as decision trees, neural networks, support vector machines, Bayesian networks, and ensemble learning are applied to evaluate different feature subset schemes. In the data mining process, the prediction performance of models is evaluated by using the repeated cross-validation method, in which the inner loop is run on ten subsamples, and the outer loop is also repeated ten times. The parameterizations involved in various methods are shown below.

5.0.1 Decision tree

The basic algorithm used in decision tree methods is the greedy algorithm, which refers to the recursive method from top to bottom. Generally, this algorithm can be divided into two steps, which involve construction of the trees and pruning them [24]. The latter step is mainly used to remove unnecessary branches and solve the overfitting problem. Four algorithms, including C4.5, Logistic Model

Table 1 Parameters of data mining for water-flooded layer evaluation

Formation	Total parameters	Parameter type	Parameter description
Lower Karamay Formation	23	Feature parameters of water-flooded layers	Water production rate (f_w), oil saturation (S_o), recovery index (F_{ow})
		Nine traditional well logs	Spontaneous potential log (SP), gamma ray log (GR), acoustic travel time log (AC), compensation density log (DEN), compensation neutron log (CNL), flush zone resistivity log (R_{xo}), transition zone resistivity log (R_i), undisturbed resistivity log (R_t)
		Lithology parameters	Median grain diameter (M_d), shale content (V_{sh}), gravel content (V_{gr})
		Physical parameters	Porosity (φ), permeability (K)
		Fluid property parameters	Irreducible water saturation (S_{wi}), residual oil saturation (S_{xo}), water saturation (S_w)
		Resistivity parameters	Difference between undisturbed resistivity and transition zone resistivity ($R_t - R_i$), difference between undisturbed resistivity and flush zone resistivity ($R_t - R_{xo}$)
		Other parameters	Formation (*ZONE*), formation thickness (H)

Tree (LMT), Classification And Regression TreeClassification And Regression Tree (CART), and Chi-Square Automatic Interaction Detection (CHAID), are applied in this study, and their parameterizations are used to solve the overfitting problem in model building.

1. The first algorithm used is C4.5. The information gain rate is utilized to overcome the drawback that information gain favors multi-value attributes, and it can address continuous-value attributes and missing data. The confidence coefficient of the pruning is 25 %.
2. The second algorithm used is LMT. Each leaf node of the tree generated by this algorithm is a logistic regression model. The classification criterion is the information gain, and cross verification is used to decide the number of loops.
3. The third algorithm used is CART. The cross verification of ten stratifications in the inner part and minimal-cost complexity pruning are both utilized to identify different feature subset schemes.
4. The fourth algorithm used is CHAID. In this algorithm, χ^2 statistical verification is used to identify optimal

Table 2 Feature subset schemes of water-flooded layer evaluation

No.	Code	Feature subset schemes	Remark
1	ALL	All 23 parameters	
2	RG1	F_{ow}, f_w, S_o	Artificially determined
3	RG2	R_t, $R_t - R_{xo}$, $R_t - R_i$, DEN, AC, CNL, φ, K, SP, GR	Artificially determined
4	RG3	F_{ow}, f_w, S_o, R_t, R_i, R_{xo}, $R_t - R_{xo}$, $R_t - R_i$, CNL, SP, ZONE	Artificially determined
5	YC	F_{ow}, f_w, S_o, ZONE, $R_t - R_{xo}$, $R - R_i$, CNL, SP	Filter: genetic algorithm for searching, LVF attribute subset evaluation
6	JQ	F_{ow}, f_w, S_o, ZONE, $R_t - R_{xo}$, $R_t - R_i$, R_t, R_{xo}, R_i, SP, φ, AC	Feature-weighted method: single-factor variance analysis

classification points and generate multi-branch trees. Its confidence factor of pessimistic error estimation is 25 %.

5.0.2 Neural networks

To identify parameter settings using neural networks, the input values are normalized to the interval [−1,1], and the stratification attributes are transformed into eight bivariate attributes. Two types of networks are designed, Back Propagation (BP) network and Radial Basis Function Network (RBFN).

1. In a BP network, a multi-layer feed-forward neural network that uses the gradient descent momentum BP algorithm is applied. This algorithm has a three-layer network structure. The number of implicit nodes is equal to half of the total number of input and output nodes, and a sigmoid function is used as the transfer function. The learning rate is set to 0.3, and the momentum is set to 0.2. In addition, 20 % of the sample data are selected as the testing set to avoid the overfitting problem, and the early termination strategy is used to terminate the training process.
2. In an RBFN network, K-means clustering analysis is used to determine the center and width of the radial basis function and the number of clusters is set to 5. The logistic regression model is also utilized to determine the weight values of implicit output.

5.0.3 Support vector machines

The preprocessing of the input attributes and input values in support vector machines is the same as that used in neural networks. Two models, C-SVC and v-SVC, are utilized, and SMO is selected as the solving algorithm. The kernel function of the two models is the radial basis function (1). To avoid overfitting, a genetic algorithm is used to choose the optimal penalty factor (C), kernel parameter (γ), and v value for each feature subset scheme. The genetic algorithm is set up as follows: the chromosome is coded using real numbers, the maximum number of generations is set to 20, the group size is set to 40, and the crossing probability is set to 0.85. For the genetic algorithm, the competing selection operator and Gaussian mutation operator are both utilized.

$$k\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right) \tag{1}$$

where x_i and x_j indicate the linear sample collection and γ is the kernel parameter, which is dimensionless.

5.0.4 Bayesian networks

Bayesian network models construct directed acyclic graphs whose nodes have conditional probability distributions, and they describe the joint probability distributions between variables [29]. Through analyzing mutual relationships among different variables, BN models utilize features of both learning and statistical inference reflected by Bayes' theorem to complete data mining tasks such as classification and clustering. Two algorithms are used in data mining with BN.

1. The first of these methods uses the simple Naive Bayes (NB) algorithm.
2. The second method uses the Trust Negotiation Algorithm (TNA) algorithm. First, all the attributes except formation are discretized, and then the TNA learning strategy and traditional Bayesian methods are both used to correct the Markov chain for the network and directly estimate conditional probabilities from the data.

5.0.5 Ensemble learning

Ensemble learning uses the AdaBoost algorithm, which is a representative lifting combination classification algorithm [32]. The AdaBoost algorithm automatically adjusts the weight values of each sample each time the lifting iteration finishes, and focuses the basis separator on incorrectly classified samples. The number of iterations is chosen to be equal to 10. Two schemes, AC5.0 and AdaBoost plus BP (ABP), were used.

1. When the AC5.0 scheme is used, the weak basis separator utilizes the C5.0 algorithm.
2. When the ABP scheme is used, the weak basis separator utilizes BP neural networks, and the parameterization is the same as in the BP scheme.

5.1 Data mining results and discussion

The open-source data mining tools Weka and RapidMiner [43] are jointly selected as the research platform to complete the data mining classification of the water-flooded layer data for the conglomerate reservoir. Based on these data mining results, the advantages and disadvantages of different algorithms and feature schemes are discussed in order to finally select the optimal prediction model for the water-flooded layer evaluation in the study area. In this study, 12 algorithms and 6 different feature subset schemes are comprehensively analyzed and compared. Figure 5 shows the results in terms of the accuracy rates from repeated cross-validation testing, and Fig. 6 shows the standard deviation of performance.

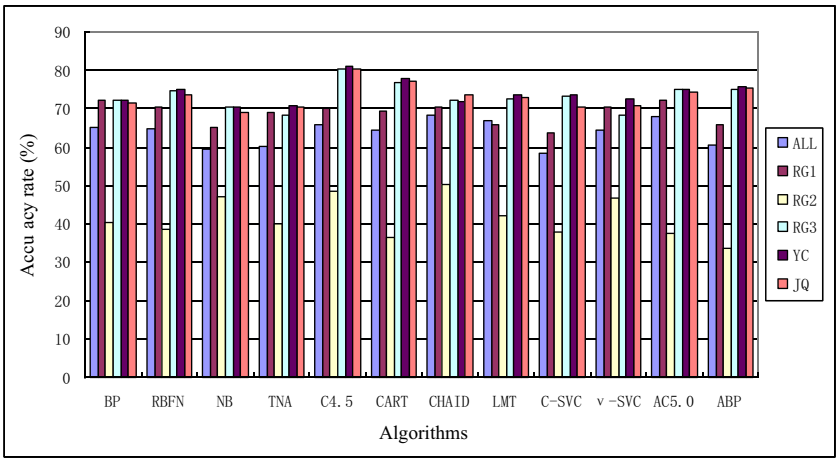


Fig. 5 Repeated cross-validation accuracy rates of various data mining algorithms

5.1.1 Comparison and decision of algorithms

Overall, the decision tree models (C4.5, LMT, CART, and CHAID) have the best performance (Fig. 5). In particular, the C4.5 and CART algorithms are superior to other algorithms (Fig. 5). The accuracy of the C4.5 algorithm reaches 81.08 % in the YC scheme, which is the best performance among all the tests. The two Bayesian algorithms (NB and TNA) have the worst performance, and their accuracies are below 70 % for all six schemes (Fig. 5). The two neural network-based methods (BP and RBFN) and the two support vector machine algorithms (C-SVC and v-SVC) have performances between those of the decision tree and Bayesian algorithms. In addition, the combination separators perform better than the independent separators, and the ensemble learning method of ABP performs well when applied to the YC scheme, where its accuracy is 75.75 %. This accuracy is higher than the accuracy of 72.34 % obtained when only the BP algorithm is applied to the YC scheme (Fig. 5).

The performance of the C4.5 algorithm is the most stable of the four decision tree methods, and its performance variance is the smallest among all 12 algorithms (Fig. 6); hence, the variations among datasets do not have a clear effect on the prediction ability of the C4.5 algorithm. Furthermore, because the differences in performance are also small among the six feature schemes when the C4.5 algorithm is used, feature scheme selection has a limited effect on the C4.5 algorithm. Based on the above analysis, considering both factors such as algorithm accuracy and the effect of variance on model evaluation, the C4.5 algorithm is selected as the final model to evaluate and identify water-flooded layers in the conglomerate reservoir of Karamay oil field.

5.1.2 Comparison and selection of feature schemes

The selection of feature schemes improves the prediction performance and ability of different algorithms, and the six feature schemes are specifically designed for data mining identification of water-flooded layers. The selection of the optimal scheme depends not only on the accuracy of the various algorithms applied to different schemes but also on the implementation convenience of the generation rules.

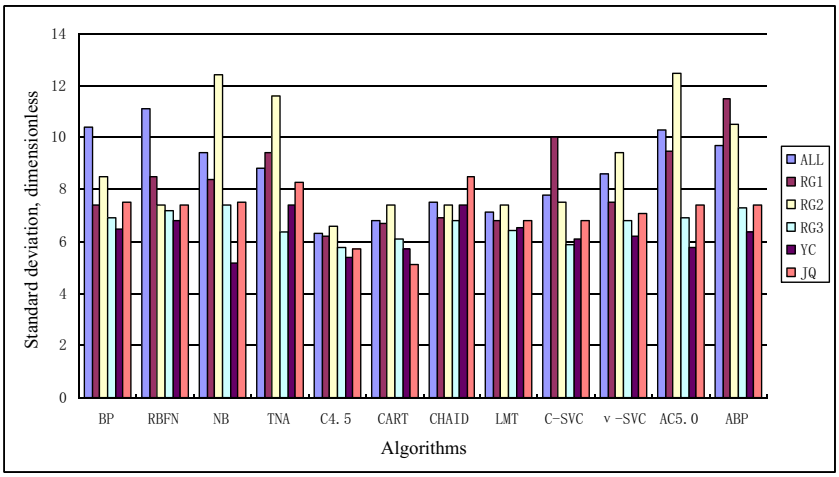


Fig. 6 Standard deviation of performance of various data mining algorithms

The C4.5 algorithm, which is the final prediction model for identifying water-flooded layers, is selected as an example to evaluate the applicability of all six feature schemes, and the corresponding results are displayed in Table 3. From the table, we can clearly see that the accuracy of the feature schemes can be generally divided into three levels, in which the accuracy of RG3, YC, and JQ are all higher than 80 % and their values are close to one another. The ALL and RG1 schemes have moderate accuracy, that is, approximately 65∼71 %. The RG2 scheme has the lowest accuracy of 48.39 %. Based on their accuracy values, the three schemes (RG3, YC, and JQ) are selected for further analysis. Considering the implementation convenience of these three schemes, it is found that the number of generation rules in the YC scheme is the smallest and its implementation convenience is also the best among these feature schemes. Therefore, the YC feature scheme as assessed using the genetic algorithm is finally selected to evaluate the water-flooded layers in the study area.

In summary, the combination of the C4.5 algorithm and the YC feature scheme obtains the best application effects for the sample datasets for the evaluation of water-flooded layers in the conglomerate reservoir, as determined based on the rule of minimum structural risk. Thus, it is selected as the final prediction model. Because decision tree methods produce "white-box" models, its mining process and results can clearly illustrate how the classifier works and the relative importance of all parameters. As shown in Table 3, the three water-flooded feature parameters (F_{ow}, f_w, and S_o) have the most significant effects on the prediction model. Among all of the six schemes, only the RG2 algorithm does not contain these three parameters, so it gives the lowest accuracy of 48.39 %; furthermore, the strong and mid-strong water-flooded layers even do not generate identification rules in the RG2 scheme (Table 3). Second, the parameter *ZONE* has a significant effect on the identification accuracy of different models. Different sublayers were deposited under different sedimentary environments and sedimentary rhythms, which influence various parameters such as lithology, physical properties, oil-bearing properties, and the pore structures of different sublayers. Therefore, layer differentiation can be used to eliminate the influences of these factors and improve model performance. In addition, from the selection of YC as the final feature scheme, we can see that the following parameters such as deep-middle resistivity difference ($R_t - R_i$), deep-shallow resistivity difference ($R_t - R_{xo}$), neutron porosity (CNL), and natural electric potential (SP) all affect the model to some extent. The resistivity difference reflects changes in resistivity when injection water enters the reservoir. Neutron porosity indicates changes in the hydrogen index of the reservoir, namely changes in the properties of fluid in the reservoir. Natural electric potential essentially reflects salinity differences between the mud filtrate and the formation water, and it provides a clear indication of the water flooding level of oil layers in its early stages. Thus, all of the four preceding parameters are moderately sensitive to the degree of flooding of layers and affect the final accuracy of the model.

6 Case study of clustering analysis

6.1 Performance evaluation of clustering analysis

Clustering analysis is an unsupervised learning algorithm, and no prior knowledge or information is used in the clustering analysis process [13]. Hence, some measures and methods need to be selected to evaluate its effectiveness and performance. The internal measurement standards of clustering analysis depend primarily on two parameters, which are the distance among observations within clusters and the distance between clusters. The former shows the compactness of the clustering results and the effectiveness of the clustering algorithm, and the latter indicates the distinguishability of the clustering algorithm as applied to specific datasets.

Table 3 The accuracy rates and generated rules of feature schemes using the C4.5 model

Feature scheme code	Accuracy rate (%)	Rule numbers				
		High flooded layer	Middle-high flooded layer	Middle flooded layer	Low flooded layer	Oil layer
ALL	65.86	5	3	6	3	4
RG1	70.17	1	1	2	1	2
RG2	48.39	0	0	3	2	1
RG3	80.29	2	1	3	2	2
YC	81.08	1	1	1	2	1
JQ	80.34	1	2	1	2	3

Table 4 Parameter statistics of reservoir clustering analysis

Formation	Total parameters	Parameter type	Parameter description
Low Karamay Formation	12	Physical parameters	Porosity, permeability
		Mercury injection parameters	Displacement pressure, retreat mercury efficiency
		Micropore structural parameters	Sorting coefficient, coefficient of variation, average value coefficient, volume ratio of pore throat, average value of capillary radius, median radius, median pressure, maximum pore throat radius

6.1.1 Internal distance in clusters

When n spatial objects are classified into K_r clusters, ($K_\mathrm{r} \in K$), where K_r is the number of clusters and K is the range of the number of clusters. The internal distance within clusters is defined as the summation of internal distances of all clusters or the summation of all distances between the objects in a particular cluster and the center of that cluster (2).

$$D = \sum_{i=1}^{K_\mathrm{r}} \sum_{h \in N_i} |h - m_i| \tag{2}$$

where D is the internal distance in the clusters (dimensionless), h is one of the spatial objects (dimensionless), and m_i is the average distance of all the objects in cluster (N_i).

6.1.2 Distance between clusters

The distance between clusters is defined as the summation of distances between the center of each cluster (that is, the average value for all objects in each cluster) and the center of all the clusters (that is, the average value for all objects in all clusters) (3).

$$L = \sum_{i=1}^{K_\mathrm{r}} |m_i - m| \tag{3}$$

where L is the distance between clusters (dimensionless), m is the average value of all objects in all clusters (dimensionless), m_i is the average value of all objects in cluster (N_i) (dimensionless), and K_r is the cluster number.

6.1.3 Performance evaluation in practice

Generally, a good clustering result will have small internal distances within clusters and large distances between clusters [34]. However, in practical applications, the relationship between the two evaluation parameters and the number of clusters must be considered comprehensively to avoid problems with overclustering and underclustering dataset. Some algorithms may result in overclustering of datasets, generating excessive numbers of clusters. Although the internal distances within clusters are small for the corresponding clustering results, the final results from the clustering algorithm are very different from the real situation, and hence, the applicability of the clustering results is relatively poor. Some algorithms may cause underclustering of datasets. Although the distance between clusters is large, datasets with the same features are not divided into more detailed clusters; thus, the value of clustering results is relatively poor in the case of underclustering.

6.2 Optimization of the clustering approach

Clustering analysis is similar to prediction modeling in that it also needs to avoid both overfitting and underfitting. Overfitting means that the model is overfitted to training datasets. When overfitting occurs, the training accuracy is relatively high but the classified results are excessively redundant and detailed, the differences between the clusters are also small, and hence, the clustering results cannot show the large-scale differences between data structures. Underfitting means that

Table 5 Internal measurement parameters of reservoir clustering results

Cluster methods	Classification of reservoir types	Distance between clusters (L)	Internal distance in clusters (D)
Algorithm based on division	5	287.53	5698.9
Algorithm based on level	10	475.3	8910.31
Algorithm based on model	3	1253.55	13,789.03
Algorithm based on density	2	1383.7	15,220.72

Table 6 Classification of conglomerate reservoir in Karamay oil field

Reservoir types	Physical parameters		Mercury injection parameters		Micropore structure parameters						
	Permeability/ $10^{-3}\,\mu\,m^2$	Porosity/%	Retreat mercury efficiency/%	Displacement pressure/MPa	Median radius/μm	Median pressure/MPa	Maximum pore throat radius/μm	Coefficient of variation	Sorting coefficient	Average coefficient	Volume ratio of pore throat	Capillary radius average/μm
Type I	1599	22.8	27.53	0.021	4.33	0.47	38.35	0.426	3.357	0.225	2.918	13.54
	376.1–2821	19.9–25.8	34.35–20.71	0.018–0.024	0.52–8.14	0.01–0.92	32.59–44.12	0.382–0.47	3.176–3.538	0.149–0.301	1.562–4.274	9.755–17.32
Type II	1284	21.3	27.61	0.018	3.81	0.37	57.78	0.452	3.468	0.198	2.68	19.01
	182.6–2386	18.2–23.4	23.79–31.43	0.01–0.026	1.51–6.11	0.02–0.72	19.69–95.87	0.391–0.513	3.271–3.665	0.167–0.229	2.166–3.194	5.851–32.16
Type III	93.93	16.4	30.16	0.131	0.99	2.12	13.68	0.287	2.742	0.179	2.981	4.414
	29.94–157.9	13.1–19.7	18.74–41.58	0.029–0.233	0.09–1.89	0.38–3.86	0.253–27.10	0.191–0.383	2.186–3.298	0.126–0.232	0.9–5.062	0.324–8.504
Type IV	56.86	18.4	43.59	0.093	0.37	2.49	13.07	0.265	2.71	0.123	1.328	2.928
	1.94–111.8	14.1–22.8	38.11–49.08	0.029–0.157	0.19–0.54	1.15–3.83	3.87–22.27	0.218–0.312	2.381–3.039	0.082–0.164	1.038–1.618	0.715–5.141
Type V	10.13	11.9	40.42	0.578	0.14	7.17	2.00	0.157	1.81	0.165	1.509	0.485
	0.503–19.75	7.93–15.9	35.15–45.69	0.277–0.879	0.06–0.23	2.71–11.6	0.163–3.841	0.122–0.192	1.496–2.124	0.126–0.204	1.178–1.84	0.096–0.874

the datasets are classified before the model has learned the true structure of the dataset. When underfitting exists, the distance between different clusters is sufficiently large and the structure is clear, but the resulting classifications have very low resolution and display large differences with the real geological objects. In that case, it is not convenient to perform detailed analysis and research. To address these two problems, the data mining workflow shown in Fig. 2 is applied to analyze and address petroleum data.

6.2.1 Collection of sample data

The data from 25 wells drilled in 2013 are selected as the basic data for reservoir type classification of the conglomerate reservoir. The basic data include parameters such as physical properties, mercury injection, and micropore structure. In total, the data contain 12 items belonging to three categories of parameters (Table 4) and 460 data records. The reservoir lithologies involved are shale sandstone, fine sandstone, conglomeratic gritstone, glutenite, and sandy conglomerate.

6.2.2 Optimal selection of clustering analysis algorithm

Four clustering analysis algorithms based on division, level, model, and density are selected for data mining classification of reservoir type using the datasets collected from the conglomerate reservoir in Karamay oil field, and both the internal distances within clusters and the distances between clusters are calculated to evaluate the performance of the classification results for each algorithm (Table 5). According to performance evaluation schemes, the reservoir types as classified by different algorithms are analyzed and the optimal clustering approach that meets the clustering classification accuracy requirements and identifies the practical features of oil reservoirs is finally selected to evaluate reservoir characteristics. As shown in Table 5, the internal distance within clusters from algorithms based on both model and density is relatively large; thus, these two algorithms have not learned the true structure of the dataset. This situation clearly illustrates that the regularity inside clusters and its compactness are both poor, and the clustering results show poor accuracy in the classification of reservoir types. The distance between clusters is relatively large (Table 5), but this result is caused by the oversimplified reservoir types as classified by both algorithms. The algorithm based on level classifies the conglomerate reservoir into ten types, and the distance between clusters is 475.3, which is larger than that from the algorithm based on division (Table 5). However, the reservoir types classified are excessively complicated, and it follows that the internal distance in clusters is thus larger than that associated with the algorithm based on division. Furthermore, the ten reservoir types from the

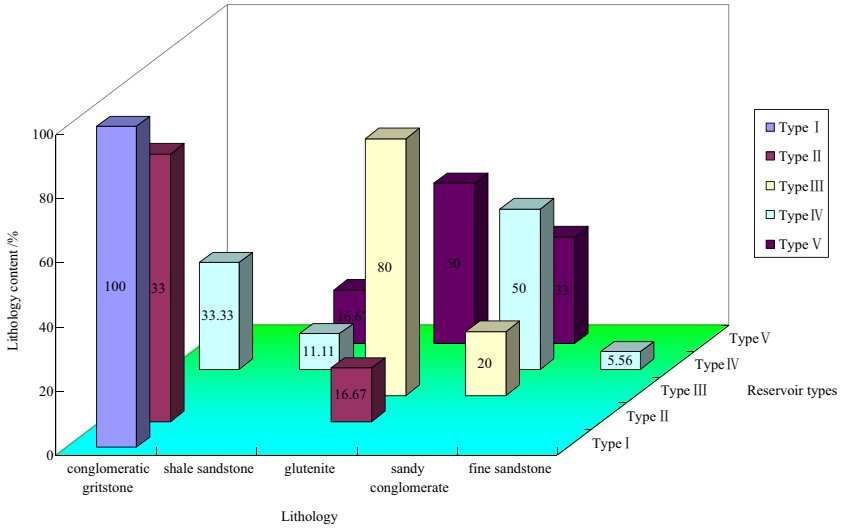


Fig. 7 Relationship between reservoir types and lithologies of conglomerate reservoir

classification results differ greatly from the lithologic types of the conglomerate reservoir; hence, it is not suitable for the comprehensive analysis of reservoir types and conglomerate lithologies. Of the four clustering models, the algorithm based on division has the smallest internal distance within clusters of 5698.9 (Table 5). It shows good compactness and regularity between clusters, and reservoir types from the classification results are also similar to the lithologic types of the conglomerate reservoir. Hence, it is suitable to apply the algorithm based on division in further research. Based on the above analysis, the algorithm based on division, also called the K-means algorithm, is selected for classification of the conglomerate reservoir types.

6.3 Analysis of clustering results

According to the 12 parameters (Table 6), the clustering algorithm based on division classifies the conglomerate

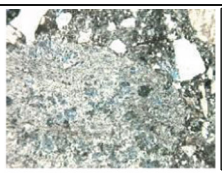
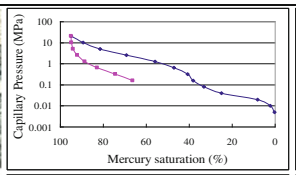
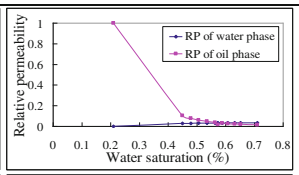
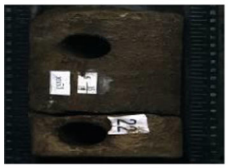
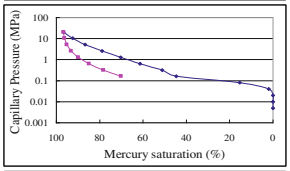
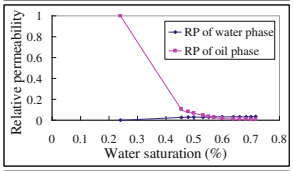
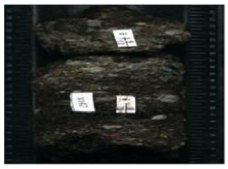
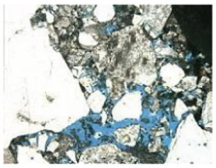
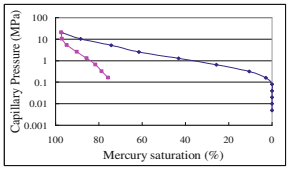
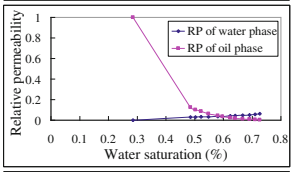
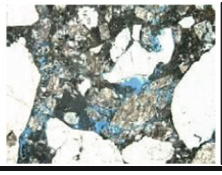
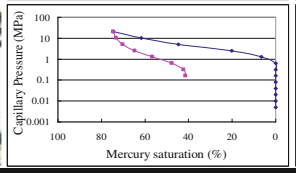
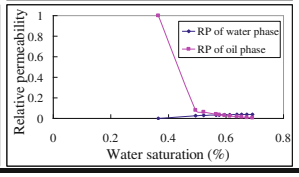
Reservoir types	Core photos	Cast thin slices	Mercury injection curves	Relative permeability curves	Water displacing oil efficiency
Type Ⅰ& Type Ⅱ					50.5%
Type Ⅲ					45.8%
Type Ⅳ					41.6%
Type Ⅴ					34.8%

Fig. 8 Comparison of micropore parameters in different reservoir types of the conglomerate reservoir

reservoir into five types, and the average, maximum, and minimum values of each parameter for each type are shown in Table 6. Generally, for sandstone reservoirs with isotropic physical properties that are not strongly heterogeneous, the lithology mainly controls reservoir type. However, because the heterogeneity of conglomerate reservoirs is relatively large, the lithology has only a moderate effect on reservoir type and is not a decisive factor. From Fig. 7, we can see that type I reservoirs contain only conglomeratic gritstone, but the other four types contain distributions of different lithologies. In particular, types IV and V contain four and three lithologies, respectively (Fig. 7). Thus, the different lithologies that make up the conglomerate reservoir in Karamay oil field might have similar physical properties and seepage properties, and any single lithology might have variable physical properties and seepage properties.

The conglomerate reservoir in the Kexia Group is classified into five types using the clustering analysis algorithm. The various properties of each reservoir type, such as core photos, thin section analysis, mercury injection curves, relative permeability curves, and the efficiency with which water displaces oil, are comparatively analyzed to evaluate their differences (Fig. 8). From type I to type V, as the reservoir features get worse, the mineral particle separation gradually becomes worse, the contact relationship between different particles becomes more complicated, and the heterogeneity of the reservoir becomes more serious (Fig. 8). In addition, from type I to type V, the mercury injection pressure changes from 0.01 to 1 MPa. Moreover, the mercury saturation of type V reservoirs only reaches 75 % (Fig. 8). These results show that from type I to type V, pore structure becomes worse, pore throat distribution becomes more heterogeneous, irreducible water saturation gradually increases from 20 to 40 %, and the final recovery of injected water correspondingly decreases from 50 to 35 % (Fig. 8). Therefore, these different reservoir types need to be accounted for when designing development programs, so that the water injection efficiency can be improved greatly for whole reservoir.

7 Conclusions and suggestions

In this study, a method for applying data mining to petroleum exploration and development is proposed to help geologists and geophysicists build reliable high-performance prediction models in a data-driven manner and, subsequently, to solve problems involving the evaluation of multiple factor-controlled, complicated geology and the development of reservoirs. The application to water-flooded layers shows that feature selection and model parameter optimization using genetic algorithms can improve model prediction accuracy. The results of the reservoir type classification show that more reasonable reservoir type classifications can be achieved in order to find some new rules for whole reservoir evaluation; if the internal measurement rules of clustering algorithms and geological features of realistic reservoirs are considered comprehensively in the data mining process, these classifications can also provide better technical support for the efficient development of oil fields. Therefore, the multiple modeling approaches provided by data mining greatly expand the library of methods for reservoir evaluation, enabling various analyses not only for predicting reservoir characteristics but also, more importantly, for finding knowledge.

However, the application of data mining methods to petroleum exploration and development is now just starting and is still in its early stages. The major task of data mining is now to effectively obtain information on geological structures, traps, reserves, and oil and gas production from massive petroleum exploration datasets. To meet this challenge, on the one hand, we need to perform additional experimental research in rock physics in order to optimize the correlation calibration between prediction models and experimental models in the data mining process, and how various factors such as datasets, features, algorithms, parameters, and model estimation methods finally affect the accuracy and operability of practical application models. On the other hand, we also need to continuously develop new data mining software and effectively integrate it with existing software such as that used in geological evaluation, well logging interpretation, and reservoir simulation in order to complete the library of algorithms used in data mining in petroleum exploration and development. Finally, the optimized solutions can be quickly provided when facing new problems.

Acknowledgments This research was supported by the Young Teacher's Research Starting Foundation of the University of Chinese Academy of Sciences (No. 55103BY00), the National Natural Science Foundation of China (No. 41574085), and the National Key Research and Development Project of China (No. 2016YFC0600310).

References

1. Aminzadeh, F.: Applications of AI and soft computing for challenging problems in the oil industry. J. Pet. Sci. Eng. **47**, 5–14 (2005)
2. Tan, F.Q., Li, H.Q., Meng, Z.X., Guo, H.F., Li, X.Y.: Research on application of data mining method in petroleum exploration and development. Oil Geophys. Prospect. **45**(1), 85–91 (2010)
3. Tan, F.Q., Li, H.Q., Liu, H.T., Jiang, F.C., Yu, H.Y.: Micro-geological causes and macro-geological controlling factors of low-resistivity oil layers in the Puao Oilfield. Pet. Sci. **6**(3), 246–253 (2009)
4. Tan, F.Q., Li, H.Q., Sun, Z.C., Min, O.Y.: Identification of natural gas fractured volcanic formation by using numerical inversion method. J. Pet. Sci. Eng. **80**(3), 85–97 (2013)

5. Li, H.Q., Guo, H.F., Guo, H.M., Meng, Z.X., Tan, F.Q., Zhang, J.: An approach of data mining for evaluation of complex formation using well logs. Acta Petrolei Sin. **30**(4), 542–549 (2009)
6. Vukelic, M.A., Miranda, E.N.: Neural networks in petroleum engineering: a case study. Int. J. Neural Syst. **7**(2), 187–194 (1996)
7. Singh, U.K., Tiwari, R.K., Singh, S.B.: One-dimensional inversion of geo-electrical resistivity sounding data using artificial neural networks—a case study. Comput. Geosci. **31**(1), 99–108 (2005)
8. Mohebbi, A., Kamalpour, R., Keyvanloo, K., Sarrafi, A.: The prediction of permeability from well logging data based on reservoir zoning, using artificial neural networks in one of an Iranian heterogeneous oil reservoir. Pet. Sci. Technol. **30**(19), 1998–2007 (2012)
9. Shi, G.R., Zhang, G.Y., Shi, X.F.: Exploration target optimization of multi-geological factors-comparative study of artificial neural network method and multiple regression analysis. Acta Petrolei Sin. **23**(5), 19–22 (2002)
10. Li, H.Q., Tan, F.Q., Xu, C.F., Wang, X.G., Peng, S.C.: Lithology identification of conglomerate reservoir base on decision tree method. J. Oil Gas Technol. **32**(3), 73–79 (2010)
11. Martinelli, G., Eidsvik, J., Sinding-Larsen, R., Rekstad, S., Mukerji, T.: Building Bayesian networks from basin-modeling scenarios for improved geological decision making. Pet. Geosci. **19**(3), 289–304 (2013)
12. Anifowose, F., Labadin, J., Abdulraheem, A.: Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines. Appl. Soft Comput. **26**(1), 483–496 (2014)
13. Han, J., Kamber, M. Data mining: concepts and techniques (the Morgan Kaufmann series in data management systems), 2nd edn. Morgan Kaufmann (2006)
14. Witten, I.H., Frank, E. Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann (2005)
15. Alimonti, C., Falcone, G.: Knowledge discovery in databases and multiphase flow metering: the integration of statistics, the data mining, neural networks, fuzzy logic, and ad hoc flow measurements towards well monitoring and diagnosis. 28–34 (2002)
16. Mohaghegh, S.D.: A new methodology for the identification of best practices in the oil and gas industry, using intelligent systems. J. Pet. Sci. Eng. **49**(3–4), 239–260 (2005)
17. Nikravesh, M.: Soft computing-based computational intelligent for reservoir characterization. Expert Syst. Appl. **26**(1), 19–38 (2004)
18. Usama, F., Gregory, P.S., Padhraic, S.: Knowledge discovery and data mining: towards a unifying framework. Portland, AAAI, 82–88 (1996)
19. Symeonidis, A.L., Mitkas, P.A.: Agent intelligence through data mining. Multi-agent systems, artificial societies, and simulated organizations series volume 14, vol. 200. International Book Series, Springer+Business Media, USA (2005)
20. Quinlan, J.: C4. 5: Programs for machine learning. Morgan Kaufmann (1993)
21. Tan, P., Steinbach, M., Kumar, V.: Introduction to data mining. Addison Wesley (2005)
22. Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, USA (1995)
23. Freedman, D.A.: Statistical models: theory and practice. Cambridge University Press (2005)
24. Quinlan, J.: Induction of decision trees. Mach. Learn. **1**, 80–112 (1986)
25. Breiman, L.: Classification and regression trees. Chapman Hall/CRC (1998)
26. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
27. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods: support vector learning table of contents, 185–208 (1999)
28. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence, 228 (1992)
29. Heckerman, D.: Bayesian networks for data mining. Data Mining Knowl Discov **1**(1), 79–119 (1997)
30. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Mach. Learn. **29**(2), 131–163 (1997)
31. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
32. Dietterich, T.G.: Ensemble methods in machine learning. Lect. Notes Comput. Sci. **18**(5), 71–15 (2000)
33. Li, S.J., Ji, H.J.: Correspondence cluster analysts and variable selection. Progress Geophys. (in Chinese) **20**(3), 694–697 (2005)
34. Xi, J.K., Tan, H.Q.: Spatial clustering analysis and its evaluation. Comput. Eng. Des. **30**(7), 1713–1715 (2009)
35. Roger, A.: Ockham's razor: a historical and philosophical analysis of Ockham's principle of parsimony (1976)
36. Elliott, S.: Let's razor Occam's razor. In: Knowles, D. (ed.) Explanation and its limits, pp. 73–93. University Press, Cambridge (1994)
37. Roald, H., Vladimir, I.M., Barry, K.C.: Ockham's razor and chemistry. HYLE–Int. J. Philos. Chem. **3**, 3–28 (1997)
38. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc., Boston (1989)
39. Marinus, E.D., Irina, O.: Connectivity of fluvial point-bar deposits: an example from the Miocene Huesca fluvial fan, Ebro Basin, Spain. AAPG Bullet. **92**(9), 1109–1129 (2008)
40. Wu, S.H., Fan, Z., Xu, C.F., Yue, D.L., Zheng, Z., Peng, S.C., Wang, W.: Internal architecture of alluvial fan the Trisaaic Lower Karamay formation in Karamay oilfield. Xinjiang. J. Paleaogeography **14**(3), 331–340 (2012)
41. Tan, F.Q., Li, H.Q., Xu, C.F., Li, Q.Y., Peng, S.C.: Quantitative evaluation methods for water-flooded layers of conglomerate reservoir based on well logging data. Pet. Sci. **7**(4), 485–493 (2010)
42. Xu, C.F., Peng, S.C., Wang, X.G., Li, H.Q., Tan, F.Q.: Quantitative interpretation of watered-out conglomerate resevior in Karamay Oil Field. Xinjiang Pet. Geol. **31**(1), 93–95 (2010)
43. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE (now: RapidMiner): rapid prototyping for complex data mining tasks. Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, 935–940 (2006)