

Optimal Prediction in Petroleum Geology by Regression and Classification Methods

Guangren Shi

Department of Expert, Research Institute of Petroleum Exploration and Development, Petro China, Beijing 100083, China

Email: grs@petrochina.com.cn

Abstract

Six methods are involved in this study: three regression methods are the regression of support vector machine (*R-SVM*), the back-propagation neural network (*BPNN*), and the multiple regression analysis (*MRA*); and three classification methods are the classification of support vector machine (*C-SVM*), the naïve Bayesian (*NBAY*), and the Bayesian successive discrimination (*BAYSD*). A proposed method optimization contains two rules: a) nonlinearity degree of a studied problem is defined by the residuals of *MRA* solution; and b) solution accuracy of a given method application is defined by the residuals of the method solution. Through eight case studies, this optimization is validated to be practical. Case studies 1 and 2 consist of both regression and classification problems, while Case studies 3~8 are classification problem. Since the regression problems of Case studies 1 and 2 have strong nonlinearity, *R-SVM*, *BPNN* and *MRA* are unavailable. However, since the classification problems of Case studies 1~8 have weak or moderate nonlinearity, *SVM* and *BAYSD* are available, whereas *NBAY* is sometimes available. Therefore, it concluded that: a) any of *R-SVM*, *BPNN* and *MRA* cannot be applied to any regression problems with strong nonlinearity, but *C-SVM*, *NBAY* or *BAYSD* could be applied if the problems are converted from regression to classification; and b) if a classification problem has weak or moderate nonlinearity, *SVM* and *BAYSD* are available, whereas *NBAY* is sometimes available.

Keywords: *Method Optimization; Nonlinearity; Solution Accuracy; Permeability; Reservoir; Trap; Fracture*

1 INTRODUCTION

In the recent years, the regression and classification methods have seen enormous success in some fields of business and sciences, but the application of these methods to petroleum geology (PG) is still in initial stage. This is because the PG is different from the other fields, with miscellaneous data types, huge quantity, different measuring precision, and lots of uncertainties to results.

Up to now, the most popular methods that have been employed in nonlinear PG problems are the following three regression methods and three classification methods [1]. The three regression methods are the regression of support vector machine (*R-SVM*), the back-propagation neural network (*BPNN*), and the multiple regression analysis (*MRA*); and three classification methods are the classification of support vector machine (*C-SVM*), the naïve Bayesian (*NBAY*), and the Bayesian successive discrimination (*BAYSD*). However, when these six methods are applied to solve a real-world problem, they often produce mixed results with varied success rates. This issue occurs due to the problem's nonlinearity degree and the different solution accuracies produced by different methods. The purpose of this paper, therefore, is how to select a proper method for real problems, i.e. a proposed method optimization. For the sake of intuitionistic views, eight case studies where the first two for the classification-regression problem and the other six for the classification problem, are demonstrated for this study.

2 REGRESSION AND CLASSIFICATION METHODS

The aforementioned regression and classification methods share the data of samples. The essential difference between the two types of methods is that the output of regression methods is real-type value and in general differs

from the real number given in the corresponding learning sample, whereas the output of classification methods is integer-type value and must be one of the integers defined in the learning samples. In the view of dataology, the integer-type value is called as discrete attribute, while the real-type value is called as continuous attribute.

The six methods (*R*-SVM, BPNN, MRA, *C*-SVM, NBAY, BAYSD) use the same known parameters, and also share the same unknown that is predicted. The only difference between them is the approach and calculation results.

Assume that there are n learning samples, each associated with $m+1$ numbers $(x_1, x_2, \dots, x_m, y^*)$ and a set of observed values $(x_{i1}, x_{i2}, \dots, x_{im}, y_i^*)$, with $i=1, 2, \dots, n$ for these numbers. In principle, $n>m$, but in actual practice $n>>m$. The n samples associated with $m+1$ numbers are defined as n vectors:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im}, y_i^*) \quad (i=1, 2, \dots, n) \quad (1)$$

where n is the number of learning samples; m is the number of independent variables in samples; $\mathbf{x}_i$ is the i^{th} learning sample vector; x_{ij} is the value of the j^{th} independent variable in the i^{th} learning sample, $j=1, 2, \dots, m$; and y_i^* is the observed value of the i^{th} learning sample.

Equation 1 is the expression of learning samples.

Let $\mathbf{x}_0$ be the general form of a vector of $(x_{i1}, x_{i2}, \dots, x_{im})$. The principles of BPNN, MRA, NBAY and BAYSD are the same, i.e. try to construct an expression, $y=y(\mathbf{x}_0)$, such that Eq. 2 is minimized. Certainly, these four different methods use different approaches and obtain calculation results in differing accuracies.

$$\sum_{i=1}^n [y(\mathbf{x}_{0i}) - y_i^*]^2 \quad (2)$$

where $y=y(\mathbf{x}_{0i})$ is the calculation result of the dependent variable in the i^{th} learning sample; and the other symbols have been defined in Eq. 1.

However, the principles of *R*-SVM and *C*-SVM methods are to try to construct an expression, $y=y(\mathbf{x}_0)$, such that to maximize the margin based on support vector points so as to obtain the optimal separating line.

This $y=y(\mathbf{x}_0)$ is called the fitting formula obtained in the learning process. The fitting formulas of different methods are different. In this paper, y is defined as a single variable.

The flowchart is as follows: the 1st step is the learning process, using n learning samples to obtain a fitting formula; the 2nd step is the learning validation, substituting n learning samples $(x_{i1}, x_{i2}, \dots, x_{im})$ into the fitting formula to get prediction values $(y_1, y_2, \dots, y_n)$, respectively, so as to verify the fitness of a method; and the 3rd step is the prediction process, substituting k prediction samples expressed with Eq. 3 into the fitting formula to get prediction values $(y_{n+1}, y_{n+2}, \dots, y_{n+k})$, respectively.

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im}) \quad (i=n+1, n+2, \dots, n+k) \quad (3)$$

where k is the number of prediction samples; $\mathbf{x}_i$ is the i^{th} prediction sample vector; and the other symbols have been defined in Eq. 1.

Equation 3 is the expression of prediction samples.

In the six methods, only MRA is a linear method whereas the other five are nonlinear methods, this is due to the fact that MRA constructs a linear function whereas the other five construct nonlinear functions, respectively.

To express the calculation accuracies of the prediction variable y for learning and prediction samples when the six methods are used, the following four types of residuals are defined.

The absolute relative residual for each sample, $R(\%)$, is defined as

$$R(\%)_i = |(y_i - y_i^*) / y_i^*| \times 100 \quad (4)$$

where y_i is the calculation result of the dependent variable in the i^{th} sample; and the other symbols have been defined in Eqs. 1 and 3. $R(\%)$ is the fitting residual to express the fitness for a sample in learning or prediction process.

It is noted that zero must not be taken as a value of y_i^* to avoid floating-point overflow. Therefore, for regression method, delete the sample if its $y_i^*=0$; and for classification method, positive integer is taken as values of y_i^* .

The mean absolute relative residual for all learning samples, $\bar{R}_1(\%)$, is defined as

$$\bar{R}_1(\%) = \sum_{i=1}^n R(\%)_i / n \quad (5)$$

where all symbols have been defined in Eqs. 1 and 4. $\bar{R}_1(\%)$ is the fitting residual to express the fitness of learning process.

The mean absolute relative residual for all prediction samples, $\bar{R}_2(\%)$, is defined as

$$\bar{R}_2(\%) = \sum_{i=n+1}^k R(\%)_i / k \quad (6)$$

where all symbols have been defined in Eqs. 3 and 4. $\bar{R}_2(\%)$ is the fitting residual to express the fitness of prediction process.

The total mean absolute relative residual for all samples, $\bar{R}^*(\%)$, is defined as

$$\bar{R}^*(\%) = \sum_{i=1}^{n+k} R(\%)_i / (n+k) \quad (7)$$

where all symbols have been defined in Eqs. 1, 3 and 4. If there are no prediction samples, $k=0$, then $\bar{R}^*(\%) = \bar{R}_1(\%)$.

$\bar{R}^*(\%)$ is the fitting residual to express the fitness of learning and prediction processes, and is adopted to express the following two rules.

2.1 Rule 1: Nonlinearity Degree of a Studied Problem

Since MRA is a linear method, its $\bar{R}^*(\%)$ for a studied problem expresses the nonlinearity degree of $y=y(\mathbf{x})$ to be solved, i.e. the nonlinearity degree of the studied problem. This nonlinearity degree can be divided into three classes: weak when $\bar{R}^*(\%) \leq 10$, moderate when $10 < \bar{R}^*(\%) \leq 30$, and strong when $\bar{R}^*(\%) > 30$.

2.2 Rule 2: Solution Accuracy of a Given Method Application

Whether linear method (MRA) or nonlinear methods (R -SVM, BPNN, C -SVM, NBAY, BAYSD), their $\bar{R}^*(\%)$ of a studied problem expresses the accuracy of $y=y(\mathbf{x})$ obtained by each method, i.e. solution accuracy of the studied problem solved by each method. This solution accuracy can be divided into three classes: high when $\bar{R}^*(\%) \leq 10$, moderate when $10 < \bar{R}^*(\%) \leq 30$, and low when $\bar{R}^*(\%) > 30$.

2.3 Methodology of the Six Methods in the Research

Through the learning process, each method constructs its own function $y=y(\mathbf{x})$. The methods of the six methods (R -SVM, BPNN, MRA, C -SVM, NBAY, BAYSD) are described below. It is noted that $y=y(\mathbf{x})$ created by BPNN is an implicit expression, i.e. which cannot be expressed as a usual mathematical formula; whereas that of the other five methods are explicit expressions, i.e. which are expressed as a usual mathematical formula.

The following will discuss the six methods. Because support vector machine (SVM) has both classification (C -SVM) and regression (R -SVM) methods, SVM is generally introduced ahead. Since the 1990's, SVM has been gradually applied in natural and social sciences, especially widely in this century. SVM is an approach utilizing machine-learning based on statistical learning theory. It is essentially performed by converting a real-world problem (the original space) into a new higher dimensional feature space using the kernel function, and then constructing a linear discriminate function in the new space to replace the nonlinear discriminate function. Theoretically, SVM can obtain the global optimal solution and avoid converging to a local optimal solution as can possibly occur in BPNN, though this problem in BPNN is rare if BPNN is properly coded [1]. The SVM procedure includes two principal methods: 1) C -SVM, such as the binary classification [1]–[6], and the ν -binary classification [7]; and 2) R -SVM, such as the ε -regression [1] [4] [6], and the ν -regression [8]. In the case studies, the binary classification for C -SVM and the ε -regression for R -SVM are employed. Moreover, it is better to take RBF (radial basis function) as a kernel function than to take the linear, polynomial and sigmoid functions under strong nonlinear conditions [1] [4], and thus the kernel function used in C -SVM and R -SVM is the RBF.

1) R -SVM

A technique of *R*-SVM, ε -regression [1] [4], has been employed. The formula created by this technique is an expression with respect to a vector $\mathbf{x}$, which is so-called a nonlinear function $y=R\text{-SVM}(x_1, x_2, \dots, x_m)$:

$$y = \sum_{i=1}^n \left[(\alpha_i^* - \alpha_i) \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2) \right] + b \quad (8)$$

where α and α^* are the vector of Lagrange multipliers, $\alpha=(\alpha_1, \alpha_2, \dots, \alpha_n)$ and $\alpha^*=(\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*)$, $0 \leq \alpha_i \leq C$ and $0 \leq \alpha_i^* \leq C$ where C is the penalty factor, and the constraint $\sum_{i=1}^n (\alpha_i^* - \alpha_i) = 0$; $\exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2)$ is a RBF kernel function; γ is the regularization parameter, $\gamma > 0$; and b is the offset of the separating hyperplane, which can be calculated using the free vectors $\mathbf{x}_i$. These free $\mathbf{x}_i$ are those vectors corresponding to $\alpha_i > 0$ and $\alpha_i^* > 0$, on which the final *R*-SVM model depends.

α_i , α_i^* , C , and γ can be solved using the dual quadratic optimization:

$$\max_{\alpha, \alpha^*} \left\{ \sum_{i=1}^n (\alpha_i^* - \alpha_i) y_i - \varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) - \frac{1}{2} \sum_{i,j=1}^n \left[(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \right] \right\} \quad (9)$$

where ε ($\varepsilon > 0$) is determined by user.

It is noted that in the case studies the formulas corresponding to Eq. 8 are not concretely written out due to their large size.

2) BPNN

The BPNN procedure has been widely applied since the 1980's [1] [9] [10], and the application of BPNN is still predominant. The formula created by BPNN is an expression with respect to m parameters ($x_1, x_2, \dots, x_m$) [1]:

$$y = BPNN(x_1, x_2, \dots, x_m) \quad (10)$$

where *BPNN* is a nonlinear function, which cannot be expressed as a usual mathematical formula and so is an implicit expression. BPNN consists of one input layer, one or more hidden layers, and one output layer. In the case studies, only one hidden layer is employed. There is no theory yet to determine how many hidden layers are needed for any given case, but in the case of output layer with only one node, it is enough to define one hidden layer. Moreover, it is also difficult to determine how many nodes a hidden layer should have. For solving local minima problem, it is suggested to use the large $N_{\text{hidden}} = 2(N_{\text{input}} + N_{\text{output}}) - 1$ estimate where N_{hidden} is the number of hidden nodes, N_{input} is the number of input nodes and N_{output} is the number of output nodes. The values of the network learning rate for the output layer and the hidden layer are within (0, 1), and in practice they can be the same.

The term back-propagation refers to the way [1] [11], the error computed at the output side is propagated backward from the output layer, to the hidden layer, and finally to the input layer. Each iteration of BPNN constitutes two sweeps: forward to calculate a solution by using a sigmoid activation function, and backward to compute the error and thus to adjust the weights and thresholds for the next iteration. This iteration is performed repeatedly until the solution agrees with the desired value within a required tolerance. The error takes the root mean square error [1] [12] is

$$RMSE(\%) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)^2} \times 100 \quad (11)$$

where y_i and y_i^* are under the conditions of normalizations in the learning process. *RMSE*(%) is used in the conditions for terminating network learning.

3) MRA

The MRA procedure has been widely applied since the 1970's [1] [13] [14], and the successive regression analysis, the most popular MRA technique, is still a very useful tool. The formula created by this technique is a linear combination with respect to m parameters ($x_1, x_2, \dots, x_m$), plus a constant term, which is so-called a linear function $y=MRA(x_1, x_2, \dots, x_m)$ [1]:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_m x_m \quad (12)$$

where the constants $b_0, b_1, b_2, \dots, b_m$ are deduced using regression criteria and calculated by the successive regression analysis of MRA. Eq. 12 is a so-called “regression equation”. In rare cases an introduced x_k can be deleted

in the regression equation, and in much rarer cases a deleted x_k could be again introduced into the regression equation. Therefore, usually Eq. 12 is solved via m iterations.

4) C-SVM

A technique of C-SVM, the binary classifier [1] [4], has been employed. The formula created by this technique is an expression with respect to a vector $\mathbf{x}$, which is so-called a nonlinear function $y=C\text{-SVM}(x_1, x_2, \dots, x_m)$:

$$y = \sum_{i=1}^n [y_i \alpha_i \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2)] + b \quad (13)$$

where α is the vector of Lagrange multipliers, $\alpha=(\alpha_1, \alpha_2, \dots, \alpha_n)$, $0 \leq \alpha_i \leq C$ where C is the penalty factor, and the constraint $\sum_{i=1}^n y_i \alpha_i = 0$; $\exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2)$ is a RBF kernel function; γ is the regularization parameter, $\gamma > 0$; and b is the offset of the separating hyperplane, which can be calculated using the free vectors $\mathbf{x}_i$. These free $\mathbf{x}_i$ are those vectors corresponding to $\alpha_i > 0$, on which the final C-SVM model depends.

α_i , C , and γ can be solved using the dual quadratic optimization:

$$\max_{\alpha} \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{2} \cdot \sum_{i,j=1}^n [\alpha_i \alpha_j y_i y_j \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)] \right\} \quad (14)$$

It is noted that in the case studies the formulas corresponding to Eq. 13 are not concretely written out due to their large size.

5) NBAY

The NBAY procedure has been widely applied since the 1990's, and widely applied in this century [1] [15]. The following introduces a NBAY technique, i.e. the naïve Bayesian. The formula created by this technique is a set of nonlinear products with respect to m parameters ($x_1, x_2, \dots, x_m$) [1] [16] [17]:

$$N_l(\mathbf{x}) = \prod_{j=1}^m \left\{ \frac{1}{\sigma_{jl} \sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu_{jl})^2}{2\sigma_{jl}^2}\right) \right\} \quad (l=1, 2, \dots, L) \quad (15)$$

where l is the class number, L is the number of classes, $N_l(\mathbf{x})$ is the discrimination function of the l^{th} class with respect to $\mathbf{x}$, σ_{jl} is the mean square error of x_j in Class l , μ_{jl} is the mean of x_j in Class l . Eq. 15 is so-called a naïve Bayesian discrimination function.

Once Eq. 15 is created, any sample shown by Eq. 1 or Eq. 3 can be substituted in Eq. 15 to obtain L values: $N_1, N_2, \dots, N_L$. If $N_{l_b} = \max_{1 \leq l \leq L} \{N_l\}$, then

$$y=l_b \quad (16)$$

for this sample.

Eq. 16 is so-called a nonlinear function $y=NBAY(x_1, x_2, \dots, x_m)$.

6) BAYSD

The BAYSD procedure has been widely applied since the 1990's, and widely applied in this century [1] [18] [19]. The following introduces BAYSD technique. The formula created by this technique is a set of nonlinear combinations with respect to m parameters ($x_1, x_2, \dots, x_m$), plus two constant terms [1]:

$$B_l(\mathbf{x}) = \ln(p_l) + c_{0l} + \sum_{j=1}^m c_{jl} x_j \quad (l=1, 2, \dots, L) \quad (17)$$

where l is the class number, L is the number of classes, $B_l(\mathbf{x})$ is the discrimination function of the l^{th} class with respect to $\mathbf{x}$, c_{jl} is the coefficient of x_j in the l^{th} discrimination function, p_l and c_{0l} are two constant terms in the l^{th} discrimination function. The constants $p_l, c_{0l}, c_{1l}, c_{2l}, \dots, c_{ml}$ are deduced using Bayesian theorem and calculated by the successive Bayesian discrimination. Eq. 17 is so-called a Bayesian discrimination function. In rare cases an introduced x_k can be deleted in the Bayesian discrimination function, and in much rarer cases a deleted x_k could be again introduced into the Bayesian discrimination function. Therefore, usually Eq. 17 is solved via m iterations.

Once Eq. 17 is created, any sample shown by Eq. 1 or Eq. 3 can be substituted in Eq. 17 to obtain L values: $B_1,$

$B_2, \dots, B_L$. If $B_{l_b} = \max_{1 \leq l \leq L} \{B_l\}$, then

$$y=l_b \quad (18)$$

for this sample.

Eq. 18 is so-called a nonlinear function $y=BAYS_D(x_1, x_2, \dots, x_m)$.

The aforementioned six methods applied to eight case studies below have the following common appoints. At the starting point of BPNN or SVM running, each feature in x and y of the learning samples and each feature in x of the prediction samples are normalized, i.e. features are normalized to $[0, 1]$ for BPNN, while features are normalized to $[-1, 1]$ for SVM (C-SVM and R-SVM). But there are no feature normalizations for samples in MRA, NBAY and BAYS_D. In order to have the comparability of results among different applications, a) for BPNN, only one hidden layer is employed, the number of hidden nodes is defined by a fixed formula with respect to the number of input nodes and the number of output nodes as described above, and the values of the network learning rate for the output layer and the hidden layer are fixed to 0.6; and b) for SVM, the termination of calculation accuracy TCA is fixed to 0.001, the RBF is taken as a kernel function, and the insensitive function ε in R-SVM is fixed to 0.1.

3 TWO CASE STUDIES OF THE CLASSIFICATION-REGRESSION PROBLEM

3.1 Case Study 1: The Permeability Prediction of an Oilfield

The objective of this case study is to calculate rock permeability and to conduct rock permeability classification (RPC) for the oil/gas wells, which has practical value when experiment data is less limited.

Using data of 12 samples from an oilfield in China, and each sample contains 3 independent variables (x_1 = irreducible water saturation, x_2 = surface area to volume ratio of rock, x_3 = porosity) and an experiment data (y^* = rock permeability), [20] adopted BPNN for the prediction of permeability. In this case study, among these 12 samples, 10 are taken as learning samples and one as prediction sample for the prediction of both permeability and RPC (Tables 1 and 2), in which for permeability by R-SVM, BPNN and MRA, and for RPC by C-SVM, NBAY and BAYS_D.

1) Regression to Calculate Rock Permeability

Using the 10 learning samples (Table 1) and by R-SVM, BPNN and MRA, the following three functions of rock permeability (y) with respect to 3 independent variables (x_1, x_2, x_3) have been constructed.

Using R-SVM, the result is an explicit nonlinear function corresponding to Eq. 8:

$$y=R-SVM(x_1, x_2, x_3) \quad (19)$$

with $C=1$, $\gamma=0.333333$, and 10 free vectors x_i .

The BPNN used consists of 3 input layer nodes, 1 output layer node and 7 hidden layer nodes. The result is an implicit nonlinear function corresponding to Eq. 10:

$$y=BPNN(x_1, x_2, x_3) \quad (20)$$

with the optimal learning time count $t_{opt}=99991$, and $RMSE(\%)=0.5156 \times 10^{-2}$.

TABLE 1 INPUT DATA FOR PERMEABILITY PREDICTION OF AN OILFIELD

Sample type	Sample No.	Three parameters related to y^a			y^*	RPC ^c
		x_1 (%)	x_2 (cm ² cm ³)	x_3 (%)	Permeability ^b (10 ⁻³ μm ²)	
Learning	1	12	1827	12.0	26	3
	2	9	1428	13.0	63	2
Samples	3	4	532	18.0	1011	1
	4	10	231	14.0	2400	1
	5	25	1169	16.0	151	2
	6	18	2310	23.7	123	2

Learning samples	7	20	2261	17.0	39	3
	8	7	644	14.0	294	1
	9	27	3542	22.0	44	3
	10	16	1253	19.7	239	1
Prediction samples	11	21	1890	17.5	(74)	(2)

^a x_1 = irreducible water saturation, x_2 = surface area to volume ratio of rock, x_3 = porosity.

^b y^* = rock permeability determined by the experimentation, number in parenthesis is not input data, but is used for calculating $R(\%)$.

^c y^* = RPC = rock permeability classification (1–high permeability, 2–intermediate permeability, 3–low permeability) determined by Table 2, number in parenthesis is not input data, but is used for calculating $R(\%)$.

TABLE 2 ROCK PERMEABILITY CLASSIFICATION BASED ON ROCK PERMEABILITY QUANTITY

Rock permeability classification	Rock permeability ($10^{-3}\mu\text{m}^2$)	RPC (rock permeability classification)
High permeability	>200	1
Intermediate permeability	[50, 200]	2
Low permeability	<50	3

Using MRA, the result is an explicit linear function corresponding to Eq. 12:

$$y=653.63+2.5543x_1-0.56583x_2+35.859x_3 \quad (21)$$

Equation 21 yields a residual variance of 0.5749 and a multiple correlation coefficient of 0.652. From the regression process, the rock permeability (y) is shown to depend on the 3 independent variables in decreasing order: x_2 , x_3 , and x_1 .

Substituting the values of 3 independent variables (x_1 , x_2 , x_3) given by the 10 learning samples and one prediction sample (Table 1) in Eqs. 19, 20 and 21, respectively, the rock permeability (y) of each sample is obtained (Table 3).

From Table 4, it shown that a) the nonlinearity degree of this studied problem is strong due to $\bar{R}^*(\%)$ of MRA is 269.7; and b) the solution accuracies of R -SVM, BPNN and MRA are low, moderate and low, respectively. Therefore, a) R -SVM and MRA are obviously unavailable; though the solution accuracy of BPNN is moderate, it is also inapplicable.

2) Classification to Calculate RPC

Using the 10 learning samples (Table 1) and by C -SVM, NBAY and BAYSD, the following three functions of RPC (y) with respect to 3 independent variables (x_1 , x_2 , x_3) have been constructed.

Using C -SVM, the result is an explicit nonlinear function corresponding to Eq. 13:

$$y=C\text{-SVM}(x_1, x_2, x_3) \quad (22)$$

with $C=128$, $\gamma=0.03125$, 8 free vectors x_i , and the cross validation accuracy CVA=72.7273%.

TABLE 3 PREDICTION RESULTS OF ROCK PERMEABILITY OF AN OILFIELD

Sample type	Sample No.	Rock permeability ($10^{-3}\mu\text{m}^2$)						
		Experimentation	Regression method					
			R -SVM		BPNN		MRA	
		y^*	y	$R(\%)$	y	$R(\%)$	y	$R(\%)$
Learning samples	1	26	137.3	428.1	26	0	80.8	210.9
	2	63	137.8	118.8	26	58.7	334.8	431.4
	3	1011	138.9	86.3	1342	32.8	1008.3	0.27
	4	2400	138.8	94.2	2400	0	1050.5	56.2
	5	151	137.6	8.9	150.5	0.36	636.0	317.1
	6	123	136.4	10.9	89.8	27.0	242.4	97.1
	7	39	136.7	250.4	26.0	33.3	34.9	10.3
	8	294	138.7	52.8	294.1	0.02	809.1	175.2
	9	44	135.7	208.5	26.0	40.9	-492.7	1220.0
	10	239	137.8	42.3	252.9	5.8	691.9	189.5
Prediction samples	11	74	137.0	85.1	26	64.9	265.4	258.6

TABLE 4 COMPARISON AMONG THE APPLICATIONS OF REGRESSION METHODS (R-SVM, BPNN AND MRA) TO ROCK PERMEABILITY PREDICTION OF AN OILFIELD

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables (x_1, x_2, x_3), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
R-SVM	Nonlinear, explicit	130.1	85.1	126.0	N/A	3 s	Low
BPNN	Nonlinear, implicit	19.9	64.9	24.0	N/A	30 s	Moderate
MRA	Linear, explicit	270.8	258.6	269.7	x_2, x_3, x_1	<1 s	Low

Using NBAY, the result is an explicit nonlinear discriminate function corresponding to Eq. 15:

$$N_l(\mathbf{x}) = \prod_{j=1}^3 \left\{ \frac{1}{\sigma_{jl} \sqrt{2\pi}} \exp \left(\frac{-(x_j - \mu_{jl})^2}{2\sigma_{jl}^2} \right) \right\} \quad (l=1, 2, 3) \quad (23)$$

where for $l=1$, $\sigma_{j1}=4.437, 371.561, 2.498, \mu_{j1}=9.25, 665, 16.425$; for $l=2$, $\sigma_{j2}=6.549, 488.409, 4.507, \mu_{j2}=17.333, 1635.667, 17.567$; for $l=3$, $\sigma_{j3}=6.128, 728.052, 4.082, \mu_{j3}=19.667, 2543.333, 17$.

Using BAYSD, the result is an explicit nonlinear discriminate function corresponding to Eq. 17:

$$\begin{cases} B_1(\mathbf{x}) = \ln(0.4) - 17.73 - 0.173x_1 - 0.015x_2 + 2.845x_3 \\ B_2(\mathbf{x}) = \ln(0.3) - 9.783 + 0.055x_1 - 0.006x_2 + 1.627x_3 \\ B_3(\mathbf{x}) = \ln(0.3) - 8.527 + 0.141x_1 + 0.004x_2 + 0.243x_3 \end{cases} \quad (24)$$

From the successive process, RPC (y) is shown to depend on the 3 independent variables in decreasing order: x_2, x_3 , and x_1 .

Though MRA is regression method but not classification method, MRA can provides the nonlinearity degree of the studied problem, and thus it is required to run MRA.

Using MRA, the result is an explicit linear function corresponding to Eq. 12:

$$y = 2.4398 + 0.01657x_1 + 0.00091x_2 - 0.12767x_3 \quad (25)$$

Equation 25 yields a residual variance of 0.10046 and a multiple correlation coefficient of 0.94844. From the regression process, RPC (y) is shown to depend on the 3 independent variables in decreasing order: x_2, x_3 , and x_1 .

Substituting the values of 3 independent variables (x_1, x_2, x_3) given by the 10 learning samples and one prediction sample (Table 1) in Eqs. 22, 23 (and then use Eq. 16), 24 (and then use Eq. 18) and 25, respectively, the RPC (y) of each sample is obtained (Table 5).

TABLE 5 PREDICTION RESULTS FROM ROCK PERMEABILITY CLASSIFICATION OF AN OILFIELD

Sample type	Sample No.	y^*	RPC ^a							
			Classification method						MRA	
			C-SVM		NBAY		BAYSD		y	R(%)
			y	R(%)	y	R(%)	y	R(%)		
Learning samples	1	3	3	0	2	33.3	3	0	2.8	7.9
	2	2	2	0	2	0	2	0	2.2	11.2
	3	1	1	0	1	0	1	0	0.7	31.0
	4	1	1	0	1	0	1	0	1.0	2.8
	5	2	2	0	2	0	2	0	1.9	6.4
	6	2	2	0	3	50	2	0	1.8	9.7
	7	3	3	0	3	0	3	0	2.7	11.7
	8	1	1	0	1	0	1	0	1.4	35.2
	9	3	3	0	3	0	3	0	3.3	9.6
	10	1	1	0	2	100	1	0	1.3	32.6
Prediction samples	11	2	2	0	2	0	2	0	2.3	13.4

^a RPC= rock permeability classification (1–high permeability, 2–intermediate permeability, 3–low permeability) determined by Table 2.

From Table 6, it shown that a) the nonlinearity degree of this studied problem is moderate due to $\bar{R}^*$ (%) of MRA is 15.6; and b) the solution accuracies of C-SVM, NBAY and BAYSD are high, moderate and high, respectively. Therefore, a) C-SVM and BAYSD are obviously available; though the solution accuracy of NBAY is moderate, it is inapplicable.

TABLE 6 COMPARISON AMONG THE APPLICATIONS OF CLASSIFICATION METHODS (C-SVM, NBAY AND BAYSD) TO ROCK PERMEABILITY CLASSIFICATION OF AN OILFIELD

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables (x_1, x_2, x_3), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
C-SVM	Nonlinear, explicit	0	0	0	N/A	5 s	High
NBAY	Nonlinear, explicit	18.3	0	16.7	N/A	<1 s	Moderate
BAYSD	Nonlinear, explicit	0	0	0	x_2, x_3, x_1	1 s	High
MRA	Linear, explicit	15.8	13.4	15.6	x_2, x_3, x_1	<1 s	Moderate nonlinearity

3.2 Case Study 2: The Oil Layer Classification of the Keshang Formation

The objective of this case study is to calculate oil layer productivity index (OLPI) and to conduct oil layer classification (OLC) in glutenite for oil wells, which has practical value when oil test data is less limited.

Using data of 14 learning samples and 4 prediction samples from the Keshang Formation in the District 8 of Kelamayi Oilfield in western China, and each sample contains 3 parameters (x_1 = porosity, x_2 = permeability, x_3 = resistivity) and an oil test result (y^* = OLPI and OLC), see Table 10.1 in [1], [1] adopted R-SVM, BPNN and MRA for the prediction of OLPI, and adopted C-SVM and BAYSD for the prediction of OLC. In this case study, NBAY is added for the prediction of OLC, and the results related to method comparison are listed in Tables 7 and 8.

4 SIX CASE STUDIES OF THE CLASSIFICATION PROBLEM

4.1 Case Study 3: The Reservoir Classification of the Fuxin Uplift

The objective of this case study is to conduct reservoir classification (RC), which has practical value when well test data is less limited.

Using data of 25 learning samples from the fourth stage of Nanquan in the Fuxin Uplift of Jilin Oilfield in eastern China, and each sample contains 5 parameters (x_1 = sandstone thickness, x_2 = porosity, x_3 = permeability, x_4 = carbonate content, x_5 = mud content) and a well test result (y^* = reservoir type), [21] adopted Bayesian discrimination for the prediction of RC. Among these 25 samples, 20 were taken as learning samples and 3 as prediction samples for the prediction of RC by C-SVM and BAYSD [1]. In this case study, among these 25 samples, 20 are taken as learning samples and one as prediction sample (Table 9) by C-SVM, NBAY and BAYSD.

TABLE 7 COMPARISON AMONG THE APPLICATIONS OF REGRESSION METHODS (R-SVM, BPNN AND MRA) TO OIL LAYER PRODUCTIVITY INDEX OF THE KESHANG FORMATION

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables (x_1, x_2, x_3), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
R-SVM	Nonlinear, explicit	209.558	867.647	355.8	N/A	3 s	Low
BPNN	Nonlinear, implicit	17.70	45.07	23.78	N/A	30 s	Moderate
MRA	Linear, explicit	81.35	414.8	155.5	x_2, x_3, x_1	<1 s	Low

TABLE 8 COMPARISON AMONG THE APPLICATIONS OF CLASSIFICATION METHODS (C-SVM, NBAY AND BAYSD) TO OIL

LAYER CLASSIFICATION OF THE KESHANG FORMATION

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables (x_1, x_2, x_3), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
C-SVM	Nonlinear, explicit	0	0	0	N/A	5 s	High
NBAY	Nonlinear, explicit	7.14	25	11.11	N/A	<1 s	Moderate
BAYSD	Nonlinear, explicit	0	0	0	x_2, x_1, x_3	1 s	High
MRA	Linear, explicit	6.26	9.59	7	x_2, x_1, x_3	<1 s	Weak nonlinearity

TABLE 9 INPUT DATA FOR RESERVOIR CLASSIFICATION OF THE FUXIN UPLIFT

Sample type	Sample No.	Well No.	Layer No.	Five parameters related to y^a					y^{*b}
				x_1 (m)	x_2 (%)	x_3 ($10^{-3}\mu\text{m}^2$)	x_4 (%)	x_5 (%)	
Learning samples	1	M-138	7	4	15.68	0.82	1.6	6.68	2
	2	M-138	5	7	14.3	0.379	3.07	16.3	3
	3	M-133	4	4.4	14.53	0.73	4.03	10.88	2
	4	M-133	4	5	16.8	1.68	2.55	7.36	2
	5	M-134	10	5.4	14.65	6	2.47	5.6	1
	6	M-134	11	2.4	15.5	10	2.8	6.2	1
	7	M-134	10	1.74	12.5	1.84	6.98	15.37	3
	8	M-134	10	2.4	13.25	1.6	1.9	13.52	3
	9	M-134-1	8	3	14.3	1.35	2.4	9.55	2
	10	M-134-1	7	1.8	13.47	0.89	2.87	16.25	3
	11	M-4	9	4.8	13.26	6.96	3.9	5	1
	12	M-4	9	1.8	13.47	4.33	5.83	15.97	3
	13	M-112	12	1.8	17	0.89	1.3	11.78	3
	14	M-112	3	3.5	9.26	0.2	2.76	18.59	3
	15	M-112	4	7	8.2	0.11	5.5	29.01	4
	16	M-112	2	2.5	6.35	0.02	8.65	25.92	3
	17	M-112	3	5	16.8	4.07	0.8	10.2	1
	18	M-135	7	3.6	11.02	0.5	1.16	14.94	3
	19	M-135	4	4.8	9.38	0.16	4.25	17	3
	20	M-135	10	5.4	8.73	0.08	2.3	13.42	3
Prediction samples	21	M-4	10	6.1	16.88	6.26	2.05	4.5	(1)

^a x_1 = sandstone thickness, x_2 = porosity, x_3 = permeability, x_4 = carbonate content, x_5 = mud content.

^b y^* = RC = reservoir classification (1–excellent, 2–good, 3–average, 4–poor) determined by the well test, number in parenthesis is not input data, but is used for calculating $R(\%)$.

Using the 20 learning samples (Table 9) and by C-SVM, NBAY and BAYSD, the following three functions of RC (y) with respect to 5 independent variables (x_1, x_2, x_3, x_4, x_5) have been constructed.

Using C-SVM, the result is an explicit nonlinear function corresponding to Eq. 13:

$$y = C\text{-SVM}(x_1, x_2, x_3, x_4, x_5) \quad (26)$$

with $C=32$, $\gamma=0.5$, 16 free vectors $\mathbf{x}_i$, and CVA=90%.

Using NBAY, the result is an explicit nonlinear discriminate function corresponding to Eq. 15:

$$N_l(\mathbf{x}) = \prod_{j=1}^5 \left\{ \frac{1}{\sigma_{jl}\sqrt{2\pi}} \exp\left(\frac{-(x_j - \mu_{jl})^2}{2\sigma_{jl}^2} \right) \right\} \quad (l=1, 2, 3, 4) \quad (27)$$

where for $l=1$, $\sigma_{j1}=1.175, 1.287, 2.142, 1.111, 2.037$, $\mu_{j1}=4.4, 15.052, 6.757, 2.493, 6.75$; for $l=2$, $\sigma_{j2}=0.728, 0.998, 0.389, 0.877, 1.683$, $\mu_{j2}=4.1, 15.327, 1.145, 2.645, 8.618$; $l=3$, $\sigma_{j3}=1.682, 2.914, 1.205, 2.325, 3.536$, $\mu_{j3}=0, 0, 0, 0, 0$; $l=4$, $\sigma_{j4}=3.1, 0.66, 0.627, 0.71, 0.71$, $\mu_{j4}=7, 8.2, 0.11, 5.5, 29.01$.

Using BAYSD, the result is an explicit nonlinear discriminate function corresponding to Eq. 17:

$$\begin{cases} B_1(\mathbf{x}) = \ln(0.2) - 95.7 + 7.581x_1 + 4.605x_2 + 9.076x_3 - 1.463x_4 + 4.599x_5 \\ B_2(\mathbf{x}) = \ln(0.2) - 67.7 + 5.235x_1 + 5.15x_2 + 3.339x_3 + 0.684x_4 + 3.408x_5 \\ B_3(\mathbf{x}) = \ln(0.55) - 76.05 + 5.169x_1 + 4.793x_2 + 4.797x_3 - 0.777x_4 + 4.763x_5 \\ B_4(\mathbf{x}) = \ln(0.05) - 152 + 8.266x_1 + 5.288x_2 + 8.53x_3 - 2.798x_4 + 7.491x_5 \end{cases} \quad (28)$$

From the successive process, RC (y) is shown to depend on the 5 independent variables in decreasing order: x_5 , x_3 , x_1 , x_4 , and x_2 .

Though MRA is regression method but not classification method, MRA can provides the nonlinearity degree of the studied problem, and thus it is required to run MRA.

Using MRA, the result is an explicit linear function corresponding to Eq. 12:

$$y = 1.895 - 0.090332x_1 + 0.004592x_2 - 0.12934x_3 - 0.043057x_4 + 0.093862x_5 \quad (29)$$

Equation 29 yields a residual variance of 0.17037 and a multiple correlation coefficient of 0.91084. From the regression process, RC (y) is shown to depend on the 5 independent variables in decreasing order: x_5 , x_3 , x_1 , x_4 , and x_2 .

Substituting the values of 5 independent variables (x_1 , x_2 , x_3 , x_4 , x_5) given by the 20 learning samples and one prediction sample (Table 9) in Eqs. 26, 27 (and then use Eq. 16), 28 (and then use Eq. 18) and 29, respectively, the RC (y) of each sample is obtained (Table 10).

From Table 11, it shown that a) the nonlinearity degree of this studied problem is moderate due to $\bar{R}^*$ (%) of MRA is 13.41; and b) the solution accuracies of C-SVM, NBAY and BAYSD are all high. Therefore, C-SVM, NBAY and BAYSD are all available.

4.2 Case Study 4: The Reservoir Classification of the Baibao Oilfield

The objective of this case study is to conduct reservoir classification (RC), which has practical value when well test data is less limited.

Using data of 29 learning samples from the Chang 3 and Chang 4+5 of Triassic Formation of Baibao Oilfield in Ordos Basin in western China, and each sample contains 4 parameters (x_1 = mud content, x_2 = porosity, x_3 = permeability, x_4 = permeability variation coefficient) and a well test result (y^* = reservoir type), [22] adopted Bayesian discrimination for the prediction of RC. Among these 29 samples, 25 were taken as learning samples and 3 as prediction samples for the prediction of RC by C-SVM and BAYSD [1]. In this case study, among these 29 samples, 24 are taken as learning samples and 3 as prediction samples for the prediction of RC (Table 12) by C-SVM, NBAY and BAYSD.

TABLE 10 PREDICTION RESULTS FROM RESERVOIR CLASSIFICATION OF THE FUXIN UPLIFT

Sample type	Sample No.	RC ^a									
		y^*	Classification method						MRA		
			C-SVM		NBAY		BAYSD				
			y	R(%)	y	R(%)	y	R(%)			y
Learning samples	1	2	2	0	2	0	2	0	2.06	2.89	
	2	3	3	0	3	0	3	0	2.68	10.8	
	3	2	2	0	2	0	2	0	2.32	15.9	
	4	2	2	0	2	0	2	0	1.88	5.79	
	5	1	1	0	1	0	1	0	1.12	11.8	
	6	1	1	0	1	0	1	0	0.92	8.26	
	7	3	3	0	3	0	3	0	2.70	10.0	
	8	3	3	0	3	0	3	0	2.72	9.36	
	9	2	2	0	2	0	2	0	2.31	15.4	
	10	3	3	0	3	0	3	0	3.08	2.70	
	11	1	1	0	1	0	1	0	0.92	7.65	
	12	3	3	0	3	0	3	0	2.48	17.3	
	13	3	3	0	3	0	3	0	2.75	8.50	
	14	3	3	0	3	0	3	0	3.22	7.39	

	15	4	4	0	3	25	4	0	3.77	5.69
	16	3	3	0	3	0	3	0	3.76	25.2
	17	1	1	0	1	0	1	0	1.92	91.7
	18	3	3	0	3	0	3	0	2.91	3.06
	19	3	3	0	3	0	3	0	2.90	3.45
	20	3	3	0	3	0	3	0	2.60	13.4
Prediction samples	21	1	1	0	1	0	1	0	0.95	5.40

^a RC = reservoir classification (1–excellent, 2–good, 3–average, 4–poor) determined by the well test.

TABLE 11 COMPARISON AMONG THE APPLICATIONS OF CLASSIFICATION METHODS (C-SVM, NBAY AND BAYSD) TO RESERVOIR CLASSIFICATION OF THE FUXIN UPLIFT

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables (x_1, x_2, x_3, x_4, x_5), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
C-SVM	Nonlinear, explicit	0	0	0	N/A	5 s	High
NBAY	Nonlinear, explicit	1.25	0	1.19	N/A	<1 s	High
BAYSD	Nonlinear, explicit	0	0	0	x_5, x_3, x_1, x_4, x_2	1 s	High
MRA	Linear, explicit	13.81	5.40	13.41	x_5, x_3, x_1, x_4, x_2	<1 s	Moderate nonlinearity

Using the 24 learning samples (Table 12) and by C-SVM, NBAY and BAYSD, the following three functions of RC (y) with respect to 4 independent variables (x_1, x_2, x_3, x_4) have been constructed.

Using C-SVM, the result is an explicit nonlinear function corresponding to Eq. 13:

$$y = C\text{-SVM}(x_1, x_2, x_3, x_4) \quad (30)$$

with $C=2048$, $\gamma=0.03125$, 8 free vectors x_i , and CVA=95.8333%.

Using NBAY, the result is an explicit nonlinear discriminate function corresponding to Eq. 15:

$$N_l(\mathbf{x}) = \prod_{j=1}^4 \left\{ \frac{1}{\sigma_{jl}\sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu_{jl})^2}{2\sigma_{jl}^2}\right) \right\} \quad (l=1, 2, 3, 4) \quad (31)$$

where for $l=1$, $\sigma_{j1}=1.551, 2.709, 2.913, 0.046$, $\mu_{j1}=10.87, 17.91, 33.252, 0.51$; for $l=2$, $\sigma_{j2}=1.617, 1.056, 4.27, 0.14$, $\mu_{j2}=13.825, 12.783, 17.917, 1.043$; for $l=3$, $\sigma_{j3}=5.562, 1.462, 0.546, 0.456$, $\mu_{j3}=17.987, 8.011, 0.85, 1.792$; for $l=4$, $\sigma_{j4}=3.1, 0.66, 0.627, 0.71$, $\mu_{j4}=15.3, 0.84, 0.652, 2.11$.

Using BAYSD, the result is an explicit nonlinear discriminate function corresponding to Eq. 17:

$$\begin{cases} B_1(\mathbf{x}) = \ln(0.167) - 81.17 + 0.559x_1 + 5.248x_2 + 1.881x_3 - 0.547x_4 \\ B_2(\mathbf{x}) = \ln(0.5) - 39.3 + 0.816x_1 + 3.871x_2 + 0.858x_3 + 2.356x_4 \\ B_3(\mathbf{x}) = \ln(0.25) - 26.4 + 1.131x_1 + 2.654x_2 - 0.312x_3 + 6.397x_4 \\ B_4(\mathbf{x}) = \ln(0.083) - 19.82 + 0.5x_1 - 0.466x_2 + 0.245x_3 + 15.275x_4 \end{cases} \quad (32)$$

From the successive process, RC (y) is shown to depend on the 4 independent variables in decreasing order: x_3, x_2, x_4 , and x_1 .

TABLE 12 INPUT DATA FOR RESERVOIR CLASSIFICATION OF THE BAIBAO OILFIELD

Sample type	Sample No.	Well No.	Layer No.	Four parameters related to y^a				y^{*b}
				x_1 (%)	x_2 (%)	x_3 ($10^{-3}\mu\text{m}^2$)	x_4	
Learning samples	1	B102	C3 ²	27.2	9.074	0.068	2.69	3
	2	B102	C3 ³	10.9	21.66	36.6	0.51	1
	3	B102	C4+5 ¹	13.74	13.64	15.9	1.01	2
	4	B107	C3 ²	18.4	0.18	0.025	2.82	4

	5	B108	C3 ³	8.48	19.27	28.62	0.57	1
	6	B108	C4+5 ¹	13.98	12.44	15.72	0.83	2
	7	B110	C4+5 ¹	10.58	12.03	14.01	1.07	2
	8	B112	C4+5 ¹	11.91	9.16	0.77	1.93	3
	9	B123	C3 ³	12.5	15.13	22.12	1.27	2
	10	B130	C3 ²	14.3	11.9	15.39	1.21	2
	11	B202	C4+5 ¹	12.2	1.5	1.28	1.4	4
	12	B205	C3 ¹	13.4	5.68	0.81	1.35	3
	13	B205	C3 ³	20.26	9.45	1.89	1.83	3
	14	B205	C4+5 ¹	12.78	12.34	12.71	1.14	2
	15	B205	C4+5 ²	12.8	14.91	33.42	0.44	1
	16	B206	C3 ³	13.26	6.35	0.98	1.58	3
	17	B206	C4+5 ¹	14.2	13.84	21.34	0.92	2
	18	B206	C4+5 ²	11.3	15.8	34.37	0.52	1
	19	B210	C3 ¹	21.89	8.35	0.58	1.37	3
	20	B210	C4+5 ¹	14.32	11.57	18.72	1.01	2
	21	B215	C3 ³	13.6	12.96	15.61	0.98	2
	22	B215	C4+5 ¹	14.03	13.78	28.23	1.0	2
	23	B217	C3 ³	17.98	11.53	20.73	0.84	2
	24	B217	C4+5 ¹	13.89	12.24	14.53	1.23	2
Prediction samples	25	B219	C3 ³	14.74	12.32	23.79	0.84	(2)
	26	H186	C4+5 ²	18.65	5.13	0.48	1.15	(3)
	27	W4	C4+5 ²	19.9	1.09	0.003	3.84	(4)

^a x_1 = mud content, x_2 = porosity, x_3 = permeability, x_4 = permeability variation coefficient.

^b y^* = RC = reservoir classification (1–excellent, 2–good, 3–average, 4–poor) determined by the well test, number in parenthesis is not input data, but is used for calculating $R(\%)$.

Though MRA is regression method but not classification method, MRA can provides the nonlinearity degree of the studied problem, and thus it is required to run MRA.

Using MRA, the result is an explicit linear function corresponding to Eq. 12:

$$y = 3.3863 - 0.003977x_1 - 0.096887x_2 - 0.020654x_3 + 0.2729x_4 \quad (33)$$

Equation 33 yields a residual variance of 0.05788 and a multiple correlation coefficient of 0.97063. From the regression process, RC (y) is shown to depend on the 4 independent variables in decreasing order: x_2 , x_3 , x_4 , and x_1 .

Substituting the values of 4 independent variables (x_1 , x_2 , x_3 , x_4) given by the 24 learning samples and 3 prediction samples (Table 12) in Eqs. 30, 31 (and then use Eq. 16), 32 (and then use Eq. 18) and 33, respectively, the RC (y) of each sample is obtained (Table 13).

From Table 14, it shown that a) the nonlinearity degree of this studied problem is weak due to $\bar{R}^*(\%)$ of MRA is 8.77; and b) the solution accuracies of C-SVM, NBAY and BAYSD are all high. Therefore, C-SVM, NBAY and BAYSD are all available.

TABLE 13 PREDICTION RESULTS FROM RESERVOIR CLASSIFICATION OF THE BAIBAO OILFIELD

Sample type	Sample No.	RC ^a								
		y^*	Classification method						MRA	
			C-SVM		NBAY		BAYSD			
			y	R(%)	y	R(%)	y	R(%)		
Learning samples	1	3	3	0	3	0	3	0	3.13	4.39
	2	1	1	0	1	0	1	0	0.63	37.2
	3	2	2	0	2	0	2	0	1.96	2.13
	4	4	4	0	4	0	4	0	4.06	1.62
	5	1	1	0	1	0	1	0	1.05	5.00
	6	2	2	0	2	0	2	0	2.03	1.36
	7	2	2	0	2	0	2	0	2.18	9.07
	8	3	3	0	3	0	3	0	2.96	1.26
	9	2	2	0	2	0	2	0	1.76	12
	10	2	2	0	2	0	2	0	2.19	9.44
	11	4	4	0	4	0	4	0	3.55	11.3
	12	3	3	0	3	0	3	0	3.13	4.48
	13	3	3	0	3	0	3	0	2.85	4.98

	14	2	2	0	2	0	2	0	2.19	9.42
	15	1	1	0	1	0	1	0	1.32	32.1
	16	3	3	0	3	0	3	0	3.13	4.31
	17	2	2	0	2	0	2	0	1.80	10
	18	1	1	0	1	0	1	0	1.24	24.3
	19	3	3	0	3	0	3	0	2.85	4.93
	20	2	2	0	2	0	2	0	2.10	4.87
	21	2	2	0	2	0	2	0	2.02	1.08
	22	2	2	0	2	0	2	0	1.69	15.7
	23	2	2	0	2	0	2	0	2.00	0.06
	24	2	2	0	2	0	2	0	2.18	9.04
Prediction samples	25	2	2	0	2	0	2	0	1.87	6.41
	26	3	3	0	3	0	3	0	3.12	3.97
	27	4	4	0	4	0	4	0	4.25	6.24

^a RC = reservoir classification (1–excellent, 2–good, 3–average, 4–poor) determined by the well test.

TABLE 14 COMPARISON AMONG THE APPLICATIONS OF CLASSIFICATION METHODS (C-SVM, NBAY AND BAYSD) TO RESERVOIR CLASSIFICATION OF THE BAIBAO OILFIELD

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables (x_1, x_2, x_3, x_4), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
C-SVM	Nonlinear, explicit	0	0	0	N/A	5 s	High
NBAY	Nonlinear, explicit	0	0	0	N/A	<1 s	High
BAYSD	Nonlinear, explicit	0	0	0	x_3, x_2, x_4, x_1	1 s	High
MRA	Linear, explicit	9.17	5.54	8.77	x_2, x_3, x_4, x_1	<1 s	Weak nonlinearity

4.3 Case Study 5: The Oil Layer Classification of the Lower H3 Formation

The objective of this case study is to conduct oil layer classification in clayey sandstone, which has practical value when oil test data is less limited.

Using data of 24 learning samples and 8 prediction samples from the lower H3 Formation in the Xia’ermen Oilfield of Henan oil province in central China, and each sample contains 8 parameters ($x_1 = SP$, $x_2 = R_{x0}1$, $x_3 = R_{x0}2$, $x_4 = \Delta t$, $x_5 = R_a0.45$, $x_6 = R_a4$, $x_7 = Cond$, $x_8 = Cal$) and an oil test result (y^* = oil layer type), see Table 10.7 in [1], [1] adopted C-SVM and BAYSD. In this case study, NBAY is added, and the results related to method comparison are listed in Table 15.

TABLE 15 COMPARISON AMONG THE APPLICATIONS OF CLASSIFICATION METHODS (C-SVM, NBAY AND BAYSD) TO OIL LAYER CLASSIFICATION OF THE LOWER H3 FORMATION

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables x_j ($j=1, 2, \dots, 8$), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
C-SVM	Nonlinear, explicit	0	0	0	N/A	5 s	High
NBAY	Nonlinear, explicit	0	0	0	N/A	<1 s	High
BAYSD	Nonlinear, explicit	0	0	0	$x_5, x_8, x_6, x_1, x_3, x_2, x_4, x_7$	1 s	High
MRA	Linear, explicit	1.68	15.11	5.04	$x_5, x_8, x_6, x_1, x_4, x_2, x_3, x_7$	<1 s	Weak nonlinearity

4.4 Case Study 6: The Trap Evaluation of the Northern Kuqa Depression

The objective of this case study is to conduct optimal selection of traps, which has practical value in the stage of rolling exploration.

Using data of 27 learning samples and 2 prediction samples from the Northern Kuqa Depression of the Tarim Basin in western China, and each sample (trap) contains 14 parameters (x_1 = unit structure, x_2 = trap type, x_3 = petroliferous formation, x_4 = trap depth, x_5 = trap relief, x_6 = trap closed area, x_7 = formation HC identifier, x_8 = data reliability, x_9 = trap coefficient, x_{10} = source rock coefficient, x_{11} = reservoir coefficient, x_{12} = preservation coefficient, x_{13} = configuration coefficient, x_{14} = resource quantity) and an integrated evaluation results by geologists (y^* = trap quality), see Table 3.6 in [1], [1] adopted C-SVM and BAYSD. In this case study, NBAY is added, and the results related to method comparison are listed in Table 16.

TABLE 16 COMPARISON AMONG THE APPLICATIONS OF CLASSIFICATION METHODS (C-SVM, NBAY AND BAYSD) TO TRAP CLASSIFICATION OF THE NORTHERN KUQA DEPRESSION

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables x_j ($j=1, 2, \dots, 14$), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
C-SVM	Nonlinear, explicit	0	0	0	N/A	5 s	High
NBAY	Nonlinear, explicit	9.26	25	10.34	N/A	<1 s	Moderate
BAYSD	Nonlinear, explicit	0	0	0	$x_{12}, x_{10}, x_9, x_4, x_5, x_{11}, x_8, x_6, x_{14}, x_1, x_2, x_7, x_{13}, x_3$	1 s	High
MRA	Linear, explicit	9.93	16.83	10.41	$x_{12}, x_{10}, x_9, x_4, x_5, x_{11}, x_8, x_6, x_{14}, x_1, x_2, x_7, x_{13}, x_3$	<1 s	Moderate nonlinearity

4.5 Case Study 7: The Fracture Prediction of Wells An1 and An2

The objective of this case study is to predict fractures using conventional well-logging data, which has practical values when the data of imaging log and core sample are limited.

Using data of 29 learning samples and 4 prediction samples from Wells An1 and An2 in the Anpeng Oilfield at the southeast of the Biyang Sag in Nanxiang Basin in central China, and each sample contains 7 parameters (x_1 = acoustictime, x_2 = compensated neutron density, x_3 = compensated neutron porosity, x_4 = micro-spherically focused resistivity, x_5 = deep laterolog resistivity, x_6 = shallow laterolog resistivity, x_7 = absolute difference of x_5 and x_6) and the imaging log (y^* = fracture identification), see Table 3.10 in [1], [1] adopted C-SVM. In this case study, NBAY and BAYSD are added, and the results related to method comparison are listed in Table 17.

4.6 Case Study 8: The Gas Layer Classification of the Tabamiao Area

The objective of this case study is to conduct gas layer classification in tight sandstones, which has practical value when gas test data is less limited.

Using data of 38 learning samples and 2 prediction samples from the Tabamiao area at the northeast of the Yishan Slope in Ordos Basin in western China, and each sample contains 3 parameters (x_1 = porosity, x_2 = permeability, x_3 = gas saturation) and a gas test result (y^* = gas layer type), see Table 4.1 in [1], [1] adopted C-SVM. In this case study, NBAY and BAYSD are added, and the results related to method comparison are listed in Table 18.

TABLE 17 COMPARISON AMONG THE APPLICATIONS OF CLASSIFICATION METHODS (C-SVM, NBAY AND BAYSD) TO FRACTURE PREDICTION OF WELLS AN1 AND AN2

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables x_j ($j=1, 2, \dots, 7$), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
C-SVM	Nonlinear, explicit	0	0	0	N/A	5 s	High
NBAY	Nonlinear, explicit	12.07	0	10.61	N/A	<1 s	Moderate
BAYSD	Nonlinear, explicit	0	0	0	$x_5, x_2, x_4, x_7, x_3, x_6, x_1$	1 s	High
MRA	Linear, explicit	16.2	22.79	16.99	$x_5, x_2, x_4, x_7, x_3, x_6, x_1$	<1 s	Moderate nonlinearity

TABLE 18 COMPARISON AMONG THE APPLICATIONS OF CLASSIFICATION METHODS (C-SVM, NBAY AND BAYSD) TO GAS LAYER CLASSIFICATION OF THE TABAMIAO AREA

Method	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on independent variables (x_1, x_2, x_3), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}1(\%)$	$\bar{R}2(\%)$	$\bar{R}^*(\%)$			
C-SVM	Nonlinear, explicit	0	0	0	N/A	5 s	High
NBAY	Nonlinear, explicit	31.14	0	29.58	N/A	<1 s	Moderate
BAYSD	Nonlinear, explicit	8.11	0	7.71	x_3, x_1, x_2	1 s	High
MRA	Linear, explicit	29.32	42.59	29.99	x_3, x_1, x_2	<1 s	Moderate nonlinearity

5 SUMMARY OF EIGHT CASE STUDIES

From Tables 4, 6, 7, 8, 11, 14, 15, 16, 17 and 18, Tables 19 and 20 summarize the two rules of the proposed method optimization applied to the eight case studies. Table 19 is for the classification-regression problem, while Table 20 is for the classification problem.

From Table 19, it is seen that if a regression problem has strong nonlinearity, it is required to convert the problem from regression to classification.

From Tables 19 and 20, it is seen that if a classification problem is of weak or moderate nonlinearity, SVM and BAYSD are available, whereas NBAY is sometimes available.

Among the aforementioned six methods, each of MRA and BAYSD can serve as a pioneering dimension-reduction tool, respectively, because the two methods all can give the dependence of the predicted value (y) on independent variables ($x_1, x_2, \dots, x_m$), in decreasing order. However, because MRA belongs to data analysis in linear correlation whereas BAYSD is in nonlinear correlation, in applications the preferable tool is BAYSD, and the next is MRA. The called “pioneering tool” is whether it succeeds or not needs a high-class nonlinear tool (e.g. BPNN for regression problem, C-SVM for classification problem) for the validation, so as to determine how many independent variables can be reduced. For instance, [6] indicated that a 16-D problem ($x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, y$) can be reduced to 8-D problem ($x_1, x_4, x_5, x_8, x_{10}, x_{11}, x_{15}, y$).

Comparing to C-SVM and BAYSD when the nonlinearity degree of a studied problem is weak or moderate, they are all available, but a) BAYSD runs much faster than C-SVM, b) it is easy to code the BAYSD program whereas very complicated to code the C-SVM program, and c) BAYSD can serve as a pioneering dimension-reduction tool, so BAYSD is conditionally better than C-SVM.

TABLE 19 SUMMARY OF CASE STUDIES 1 AND 2

Case study	Nonlinearity defined by Rule 1	Regression method			Classification method		
		R-SVM	BPNN	MRA	C-SVM	NBAY	BAYSD
Case study 1 (regression) See Table 4	Strong	Unavailable ($\bar{R}^*(\%)=126.0$)	Unavailable $\bar{R}^*(\%)=24.0$	Unavailable ($\bar{R}^*(\%)=269.7$)	N/A	N/A	N/A
Case study 1 (classification) See Table 6	Moderate	N/A	N/A	$\bar{R}^*(\%)=15.6$	Available ($\bar{R}^*(\%)=0$)	Unavailable ($\bar{R}^*(\%)=16.7$)	Available ($\bar{R}^*(\%)=0$)
Case study 2 (regression) See Table 7	Strong	Unavailable ($\bar{R}^*(\%)=355.8$)	Unavailable $\bar{R}^*(\%)=23.78$	Unavailable ($\bar{R}^*(\%)=155.5$)	N/A	N/A	N/A
Case study 2 (classification) See Table 8	Weak	N/A	N/A	$\bar{R}^*(\%)=7$	Available ($\bar{R}^*(\%)=0$)	Unavailable ($\bar{R}^*(\%)=11.11$)	Available ($\bar{R}^*(\%)=0$)

TABLE 20 SUMMARY OF CASE STUDIES 3 TO 10

Case study	Nonlinearity	MRA	Classification method
------------	--------------	-----	-----------------------

	defined by Rule 1		C-SVM	NBAY	BAYSD
Case study 3 (classification) See Table 11	Moderate	$\bar{R}^*(\%) = 13.41$	Available ($\bar{R}^*(\%) = 0$)	Available ($\bar{R}^*(\%) = 1.19$)	Available ($\bar{R}^*(\%) = 0$)
Case study 4 (classification) See Table 14	Weak	$\bar{R}^*(\%) = 8.77$	Available ($\bar{R}^*(\%) = 0$)	Available ($\bar{R}^*(\%) = 0$)	Available ($\bar{R}^*(\%) = 0$)
Case study 5 (classification) See Table 15	Weak	$\bar{R}^*(\%) = 5.04$	Available ($\bar{R}^*(\%) = 0$)	Available ($\bar{R}^*(\%) = 0$)	Available ($\bar{R}^*(\%) = 0$)
Case study 6 (classification) See Table 16	Moderate	$\bar{R}^*(\%) = 10.41$	Available ($\bar{R}^*(\%) = 0$)	Unavailable ($\bar{R}^*(\%) = 10.34$)	Available ($\bar{R}^*(\%) = 0$)
Case study 7 (classification) See Table 17	Moderate	$\bar{R}^*(\%) = 16.99$	Available ($\bar{R}^*(\%) = 0$)	Unavailable ($\bar{R}^*(\%) = 10.61$)	Available ($\bar{R}^*(\%) = 0$)
Case study 8 (classification) See Table 18	Moderate	$\bar{R}^*(\%) = 29.99$	Available ($\bar{R}^*(\%) = 0$)	Unavailable ($\bar{R}^*(\%) = 29.58$)	Available ($\bar{R}^*(\%) = 7.71$)

6 CONCLUSIONS

In this paper, a method optimization is proposed. Through the aforementioned eight case studies, five major conclusions can be drawn as follows:

- 1) the proposed two rules (nonlinearity degree of a studied problem, solution accuracy of a given method application) of the proposed method optimization are practical;
- 2) the total mean absolute relative residual $\bar{R}^*(\%)$ of MRA can be used to measure the nonlinearity degree of a studied problem, and thus MRA should be run at first;
- 3) any of *R*-SVM, BPNN and MRA cannot be applied to any regression problems with strong nonlinearity, but C-SVM, NBAY or BAYSD could be applied if the problems are converted from regression to classification;
- 4) if a classification problem has weak or moderate nonlinearity, SVM and BAYSD are available, whereas NBAY is sometimes available;
- 5) in general, the preferable method for regression problems is BPNN, and the next are *R*-SVM and MRA, while the preferable method for classification problems with weak or moderate nonlinearity is C-SVM or BAYSD, and the next is NBAY.

ACKNOWLEDGMENT

This work was supported by the Research Institute of Petroleum Exploration and Development (RIPED) and PetroChina.

REFERENCES

- [1] Shi G. "Data Mining and Knowledge Discovery for Geoscientists." Elsevier Inc, USA, 2013
- [2] Shi G. "The use of support vector machine for oil and gas identification in low-porosity and low-permeability reservoirs." Int J Math Model Numer Optimisa 1(1/2): 75-87, 2009
- [3] Shi G, Yang X. "Optimization and data mining for fracture prediction in geosciences." Procedia Comput Sci 1(1): 1353-1360, 2010
- [4] Chang C, Lin C. "LIBSVM: a library for support vector machines, Version 3.1." Retrieved from www.csie.ntu.edu.tw/~cjlin/libsvm, 2011
- [5] Zhu Y, Shi G. "Identification of lithologic characteristics of volcanic rocks by support vector machine." Acta Petrolei Sinica 34(2): 312-322, 2013
- [6] Shi G, Zhu Y, Mi S, Ma J, Wan J. "A big data mining in petroleum exploration and development." Adv Petrol Expl Devel 7(2): 1-8, 2014

- [7] Chang C, Lin C. "Training ν -support vector classifiers: Theory and algorithms." *Neural Comput* 13(9): 2119-2147, 2001
- [8] Chang C, Lin C. "Training ν -support vector regression: Theory and algorithms." *Neural Comput* 14(8): 1959-1977, 2002
- [9] Darabi H, Kavousi A, Moraveji M, Masihi M. "3D fracture modeling in Parsi oil field using artificial intelligence tools." *J Petro Sci Eng* 71(1-2): 67-76, 2010
- [10] Labani MM, Kadhodaie-Ilkhchi A, Salahshoor K. "Estimation of NMR log parameters from conventional well log data using a committee machine with intelligent systems: A case study from the Iranian part of the South Pars gas field." *Persian Gulf Basin J Petro Sci Eng* 72(1-2): 175-185, 2010
- [11] Güler İ, Übeyli ED. "Detection of ophthalmic artery stenosis by least-mean squares backpropagation neural network." *Comput Biol Med* 33(4): 333-343, 2003
- [12] Hush D R, Horne B G. "Progress in supervised neural networks." *IEEE Sig Proc Mag* 10(1): 8-39, 1993
- [13] Sharma MSR, O'Regan M, Baxter CDP, Moran K, Vaziri H, Narayanasamy R. "Empirical relationship between strength and geophysical properties for weakly cemented formations." *J Petro Sci Eng* 72(1-2): 134-142, 2010
- [14] Singh J, Shaik B, Singh S, Agrawal VK, Khadikar PV, Deeb O, Supuran CT. "Comparative QSAR study on para-substituted aromatic sulphonamides as CAII inhibitors: information versus topological (distance-based and connectivity) indices." *Chem Biol Drug Design* 71, 244-259, 2008
- [15] Ramoni M, Sebastiani P. "Robust Bayes classifiers." *Artificial Intelligence* 125(1-2): 207-224, 2001
- [16] Tan P, Steinbach M, Kumar V. "Introduction to Data Mining." Pearson Education, Boston, MA, USA, 2005
- [17] Han J, Kamber M. "Data Mining: Concepts and Techniques, 2nd ed." Morgan Kaufmann, San Francisco, CA, USA, 2006
- [18] Denison DGT, Holmes CC, Mallick BK, Smith AFM. "Bayesian Methods for Nonlinear Classification and Regression." John Wiley & Sons Inc, Chichester, England, UK, 2002
- [19] Shi G. "Four classifiers used in data mining and knowledge discovery for petroleum exploration and development." *Adv Petrol Expl Devel* 2(2): 12-23, 2011
- [20] Yang J, Yang C, Zhang Y Cui L, Wang L. "Permeability prediction method based on improved neural network." *Lithologic Reservoirs* 23(1): 98-102, 2011
- [21] Fu D, Xu J, Wang G. "Reservoir classification and evaluation based on Q cluster analysis combined with Bayesian discrimination algorithm." *Sci Tech Review*, 29(3): 29-33, 2011
- [22] Zhang W, Li X, Jia G. "Quantitative classification and evaluation of low permeability reservoir, with Chang 3 and Chang 4+5 of Triassic Formation in Baibao Oil Field as examples." *Sci Tech Review*, 26(21): 61-65, 2008

AUTHOR



Guangren Shi was born in Shanghai, China in February, 1940, Professor with qualification of directing Ph. D. students. He was graduated from Xi'an Jiaotong University, China in 1963, majoring in applied mathematics (1958–1963). Since 1963, he has been engaged in computer applications for petroleum exploration and development. In the recent 30 years, his research only contains two fields of basin modeling (petroleum system) and data mining for geosciences. In the recent years, however, he has focused on the latter more than the former.

He has more than 50 years of professional career, working for the computer center, Daqing Oilfield, petroleum ministry, China as Associate Engineer and Head of software group (1963–1967), the computer center, Shengli Oilfield, petroleum ministry, China as Engineer and Director (1967–1978), the computer center, petroleum ministry, China as Engineer and Head of software group (1978–1985), the Aldridge laboratory of applied geophysics, Columbia university at New York City, U.S.A as Visiting Scholar (1985–1987), the computer application technology research department, Research Institute of Petroleum Exploration and Development (RIPED), China National Petroleum Corporation (CNPC), China as Professor, Director (1987–1997), the RIPED, PetroChina Company Limited (PetroChina), China as Professor with qualification of directing Ph. D. students, Deputy Chief Engineer (1997–2001), and the department of specialists in RIPED of PetroChina, China as Professor with qualification of directing Ph. D. students (2001–Present). He published eight books, in which there are three in English, i.e. 1) Shi G. R. 2013. *Data Mining and Knowledge Discovery for Geoscientists*. Elsevier Inc, USA. 367pp; 2) Shi G. R. 2005. *Numerical Methods of Petroliferous Basin Modeling*, 3rd edition. Petroleum Industry Press, Beijing, China. 338pp, which was book-reviewed by Mathematical Geosciences in 2009; and 3) Shi G. R. 2000. *Numerical Methods of Petroliferous Basin Modeling*, 2nd edition. Petroleum Industry Press, Beijing, China. 233pp, which was book-reviewed by Mathematical Geology in 2006. He also published 74 articles, in which there are 16 in English, e.g. three articles indexed by SCI: 1) Shi G. R. 2009. A simplified dissolution-precipitation

model of the smectite to illite transformation and its application. *Journal of Geophysical Research-Solid Earth*, 114, B10205, doi:10.1029/2009JB006406; 2) Shi G. R. 2008. Basin modeling in the Kuqa Depression of the Tarim Basin (Western China): A fully temperature-dependent model of overpressure history. *Mathematical Geosciences*, 40 (1): 47–62; and 3) Shi G. R., Zhou X. X., Zhang G. Y., Shi X. F., Li H. H. 2004. The use of artificial neural network analysis and multiple regression for trap quality evaluation: a case study of the Northern Kuqa Depression of Tarim Basin in western China. *Marine and Petroleum Geology*, 21 (3): 411–420.

Prof. Shi is Member of Society of Petroleum Engineers (International), Member of Chinese Association of Science and Technology, Member of Petroleum Society of China. And he also is Regional Editor (Asia), International Journal of Mathematical Modelling and Numerical Optimisation, and Member of Editorial Board, Journal of Petroleum Science Research. He received three honors: 1) A Person Studying Overseas and Return with Excellent Contribution, appointed by the Ministry of Education of China (1991); 2) Special Government Allowance, awarded by the State Council of China (1994); and 3) Grand Award of Sun Yueqi Energy, awarded by the Ministry of Science-Technology of China (1997). And he also obtained four awards of Science-Technology Progress, in which one is China National Award, and three are from CNPC and PetroChina.