CSCanada
Advances in Petroleum Exploration and Development
Vol. 2, No. 2, 2011, pp. 12-23
DOI:10.3968/ j.aped.1925543820110202.107

ISSN 1925-542X[Print]
ISSN 1925-5438[Online]
www.cscanada.net
www.cscanada.org

Four Classifiers Used in Data Mining and Knowledge Discovery for Petroleum Exploration and Development

SHI Guangren[1,*]

[1]Research Institute of Petroleum Exploration and Development, PetroChina, P. O. Box 910, Beijing, 100083, China.
[*]Corresponding author.
Email: grs@petrochina.com.cn

Supported by the Research Institute of Petroleum Exploration and Development (RIPED) and PetroChina. Thank Dr Ben F McLean for helping to check and edit the English manuscript.

Received 12 November 2011; accepted 18 December 2011

Abstract

The application of data mining and knowledge discovery in databases for petroleum exploration and development (PE&D) is becoming promising, though still at an early stage. Up to now, the data mining tools usually used in PE&D are four classifiers: multiple regression analysis (MRA), Bayesian discrimination (BAYD), back-propagation neural network (BPNN), and support vector machine (SVM). Each of the four classifiers has its advantages and disadvantages. A question, however, has been raised in applications is: which classifier is the most applicable to a specified application? This paper has given an answer to the question through two case studies: 1) trap quality evaluation of the Northern Kuqa Depression of the Tarim Basin in western China, and 2) oil identification of the Xiefengqiao anticlinal structure of the Jianghan Basin in central China. Case 1 shows that the results of BAYD, BPNN and SVM are same and can have zero residuals, while MRA has unallowable residuals; but Case 2 shows that the results of only SVM have zero residuals, while BAYD, BPNN and MRA have unallowable residuals. The reasons are: a) since the two cases are nonlinear problems, the linear MRA is not applicable; b) since the nonlinearity of Case 1 is weak, the nonlinear BAYD, BPNN and SVM are applicable; and c) since the nonlinearity of Case 2 is strong, only nonlinear SVM is applicable. Therefore, it is proposed that: we can adopt MRA when a problem is linear; adopt BAYD, BPNN, or SVM when a problem is weakly nonlinear; and adopt only SVM when a problem is strongly nonlinear. In addition, the predictions of the applicable classifiers coincide with real exploration results, and a commercial gas trap was discovered after the forecast in Case 1 and SVM can correct some erroneous well-log interpretations in Case 2.

Key words: Multiple regression analysis; Bayesian discrimination; Back-propagation neural network; Support vector machine; Trap quality evaluation; Oil identification

Shi, G.R. (2011). Four Classifiers Used in Data Mining and Knowledge Discovery for Petroleum Exploration and Development. *Advances in Petroleum Exploration and Development, 2*(2), 12-23. Available from: URL: http://www.cscanada.net/index.php/aped/article/view/j.aped.1925543820110202.107
DOI: http://dx.doi.org/10.3968/j.aped.1925543820110202.107

INTRODUCTION

The data to be studied in data mining are divided into two categories: the learning samples, and the prediction samples. Each learning sample contains y (an object) value and $\boldsymbol{x}$ (the related parameters of y) values, whereas each prediction sample only contains $\boldsymbol{x}$ values. Data mining consists of three steps (Fig. 1): 1) "Data preprocessing", such as data selection, data cleaning, dandling missing data, identifying misclassifications, identifying outliers, data transformation, min–max normalization, etc; 2) "Knowledge discovery", using the learning samples (y, $\boldsymbol{x}$), to use a proper data mining algorithm to obtain an expression $y=f(\boldsymbol{x})$ that is so-called new knowledge discovered. 3) "Knowledge application", using the prediction samples ($\boldsymbol{x}$), to substitute each $\boldsymbol{x}$ values in $y=f(\boldsymbol{x})$, each y value is obtained.

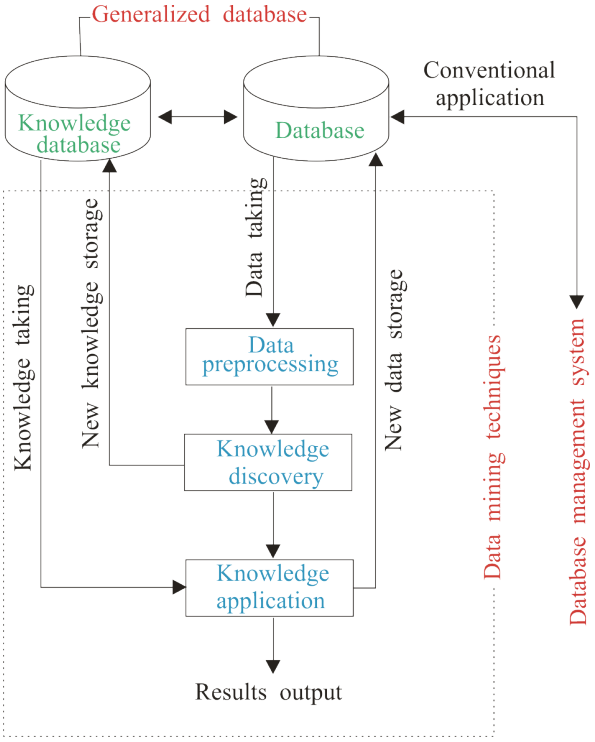

Figure 1
Flowchart of Data Mining

In the early 21 century, data mining was predicted to be "one of the most revolutionary developments of the next decade," and chosen as one of 10 emerging technologies that will change the world[1, 2, 3]. In fact, in the recent 20 years, the field of data mining has seen enormous success, both in terms of broad-ranging application achievements and in terms of scientific progress and understanding[4]. Data mining is the computerized process of extracting previously unknown and important actionable information and knowledge from large databases. This knowledge can then be used to make crucial decisions by leveraging the individual's intuition and experience to objectively generate opportunities that might otherwise go undiscovered. So Data mining is also called discovering knowledge in data (Fig. 1). It has been widely used in some fields of business and sciences, but the data mining application to petroleum exploration and development (PE&D) is still in initial stage[5, 6, 7, 8].

Facing the large amount of PE&D databases, people can use the database management system to conduct conventional applications (such as query, searches and simple statistical analysis), but cannot obtain the available knowledge inhered in data, falling in a puzzledom of 'rich data but poor knowledge'. The only solution is to develop data mining techniques (Fig. 1) in PE&D databases.

The application of data mining and knowledge discovery in databases for PE&D is becoming promising, though still at an early stage. Up to now, the data mining tools usually used in PE&D are four classifiers: multiple regression analysis (MRA), Bayesian discrimination (BAYD), back-propagation neural network (BPNN), and support vector machine (SVM). Each of the four classifiers has its advantages and disadvantages. A question, however, has been raised in applications is: which classifier is the most applicable to a specified application? This paper has given an answer to the question through two case studies below: 1) trap quality evaluation of the Northern Kuqa Depression of the Tarim Basin in western China, and 2) oil identification of the Xiefengqiao anticlinal structure of the Jianghan Basin in central China.

1. OVERVIEW OF THE FOUR CLASSIFIERS

In this study, MRA, BAYD, BPNN, and SVM use the same known parameters, and also share the same unknown to be predicted. Only the results calculated by each classifier are different (Figs. 2 and 3).

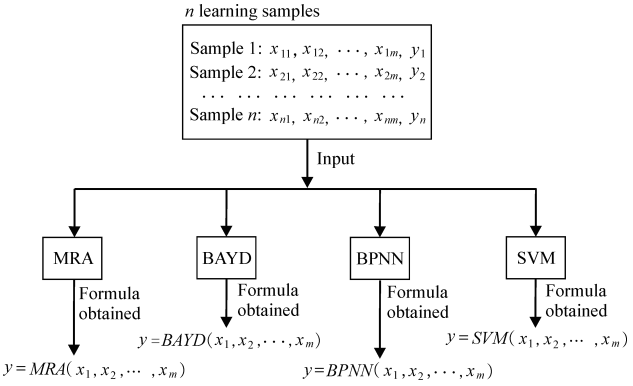

Figure 2
Sketch Map of the Learning Process for MRA, BAYD, BPNN, and SVM

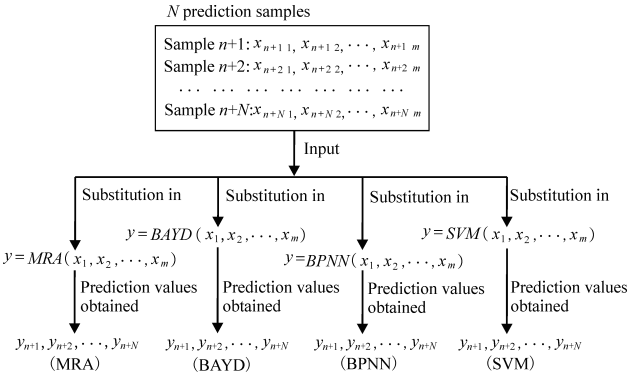

Figure 3
Sketch Map of the Prediction Process for MRA, BAYD, BPNN, and SVM

Learning process (Fig. 2)

Assume that there are n learning samples, each associated with m+1 independent parameters (x_1, x_2, …, x_m, y) and a set of observed values (x_{1i}, x_{2i}, …, x_{mi}, y_i), with i=1, 2,

…, n for these parameters. In principle, $n>m-1$, but in actual practice $n>>m-1$. The n samples associated with m parameters are defined as n vectors:

$$\boldsymbol{x}_i=(x_{1i}, x_{2i}, \ldots, x_{mi}) \quad (i=1, 2, \ldots, n) \tag{1}$$

Let $\boldsymbol{x}$ be the general form of a vector as defined in Eq. (1). The principles of MRA, BAYD, BPNN and SVM are the same, i.e. try to construct an expression, $y=f(\boldsymbol{x})$, such that

$$\sum_{i=1}^{n}\left[f\left(\boldsymbol{x}_i\right)-y_i\right]^2 \tag{2}$$

is minimized. However, in detail the different classifiers use different approaches, and provide results of differing accuracy.

Prediction process (Fig. 3)

Assume that there are N prediction samples, each associated with m independent parameters (x_1, x_2, …, x_m) and a set of observed values (x_{1i}, x_{2i}, …, x_{mi}), with $i=n+1, n+2, \ldots, n+N$ for these parameters. The N samples associated with m parameters are defined as N vectors:

$$\boldsymbol{x}_i=(x_{1i}, x_{2i}, \ldots, x_{mi}) \quad (i=n+1, n+2, \ldots, n+N) \tag{3}$$

In Figs. 2 and 3, *MRA* is a linear function, whereas *BAYD*, *BPNN* and *SVM* are nonlinear function. These characteristics will be seen in the following instructions and case studies.

1.1 MRA

The MRA procedure was established in the 1970's, and has been widely applied in the natural and social sciences[e.g. 9]. Successive regression analysis, the most popular MRA technique, is still a very useful tool in some fields[e.g. 8, 10, 11, 12, 13].

The formula created using MRA is the following linear combination with respect to m parameters (x_1, x_2, …, x_m), plus a constant term, which is so-called $y=MRA(x_1, x_2, \ldots, x_m)$ in Figs. 2 and 3:

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_mx_m \tag{4}$$

where the constants b_0, b_1, b_2, …, b_m are deduced using regression criteria and calculated by the successive regression analysis of *MRA*. Eq. (4) is a so-called "regression" equation. As indicated in Fig. 4, 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. (4) is solved via m iterations (Fig. 4).

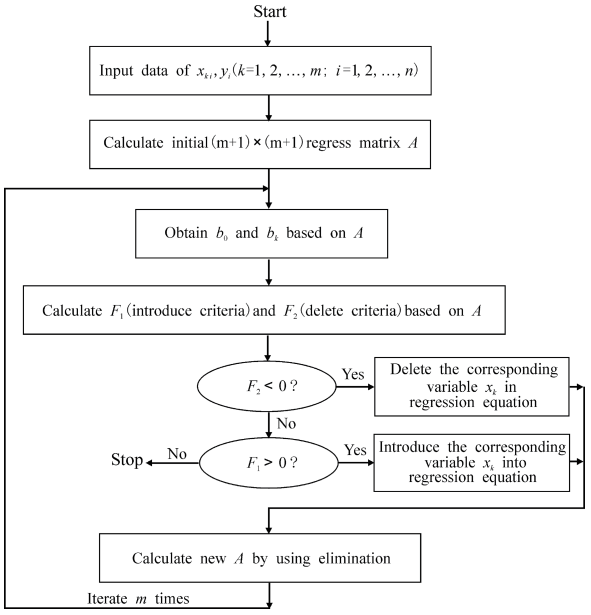

Figure 4
Sketch Map of the Learning Process by Successive Regression Analysis

1.2 BAYD

The BAYD procedure has been widely applied in the natural and social sciences since the 1990's[14]. Bayesian discrimination, the most popular Bayesian technique, is still a very useful tool in some fields[e.g. 15, 16]. The following introduces a BAYD technique: the successive Bayesian discrimination.

The formula created using BAYD is the following a set of nonlinear combinations with respect to m parameters (x_1, x_2, …, x_m), plus two constant terms:

$$B_l(\boldsymbol{x}) = \ln(p_l) + c_{0l} + \sum_{j=1}^{m} c_{jl}x_j \quad (l = 1, 2, \ldots L) \tag{5}$$

where l is the class number, L is the number of classes, $B_l(\boldsymbol{x})$ is the discrimination function of the l^{th} class of y with respect to $\boldsymbol{x}$, 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 lth discrimination function. The constants p_l, c_{0l}, c_{1l}, c_{2l}, …, c_{ml} are deduced using Bayesian theorem and calculated by the successive Bayesian discrimination of BAYD. Eq. (5) is a so-called Bayesian discrimination function. As indicated in Fig. 5, 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. (5) is solved via m iterations (Fig. 5).

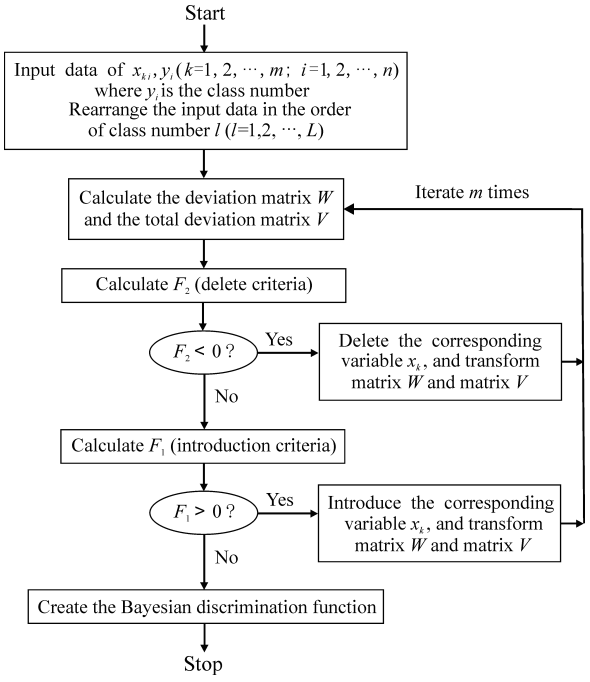

Figure 5
Sketch Map of the Learning Process by Successive Bayesian Discrimination

Once Eq. (5) is created, we can substitute a sample shown by Eq. (1) or (3) in Eq. (5) to obtain L values: B_1, B_2, …, B_L.

$$\text{If } B_{l_b}(\boldsymbol{x}) = \max_{1\le l\le L}\{B_l(\boldsymbol{x})\} \tag{6}$$

$$\text{then } y=l_b \tag{7}$$

Eq. (7) is so-called $y= BAYD(x_1, x_2, \ldots, x_m)$ in Figs. 2 and 3.

1.3 BPNN

The BPNN procedure was established in the 1980's[e.g. 17, 18], and the application of BPNN is still predominant[e.g. 11, 13, 19, 20, 21, 22].

The formula created using BPNN is the following implicit expression with respect to m parameters ($x_1, x_2, \ldots, x_m$), which is mentioned in Figs. 2 and 3:

$$y=BPNN(x_1, x_2, \ldots, x_m) \tag{8}$$

where *BPNN* is a nonlinear function constructed by error BPNN (Fig. 6), which cannot be expressed as a usual mathematical formula and is an implicit expression. Fig. 6 illustrates only one hidden layer, but in practice BPNN can have more than one hidden layer. There is no theory yet to determine how many hidden layers are needed for any given case, but in the case of only one output layer, 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 $N_{hidden}=2(N_{input}+N_{output})-1$ 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 and taken as 0.6.

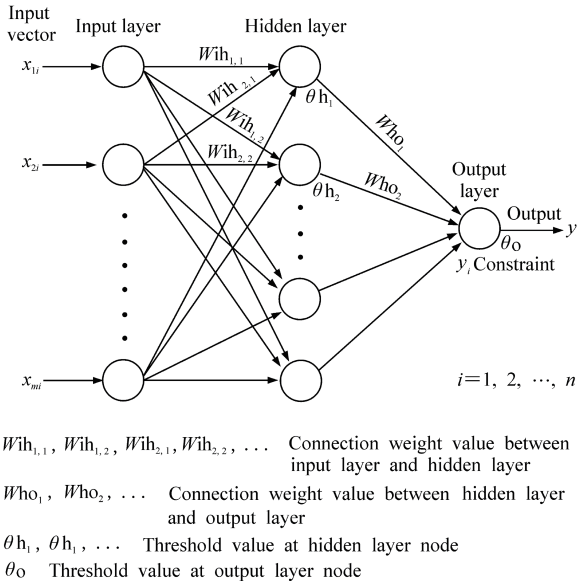

Figure 6
Sketch Map of the Learning Process by Back-Propagation Neural Network

The term back-propagation refers to the way[13, 19]: 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 $f(z)=1/[1+\exp(z)]$, and backward to compute the error and thus to adjust the weights and thresholds for the next iteration (Fig. 6). This iteration is performed repeatedly until the solution agrees with the desired value within a required tolerance.

1.4 SVM

The SVM procedure was established in the 1990's, and includes two principal methods: LP-SVM, the linear programming SVM[23]; and *C*-SVM, the binary classification SVM[24]. Numerous methods for feature selection in SVMs have been proposed[25, 26].

SVM is a new approach utilizing machine-learning based on statistical learning theory. It is essentially performed by converting a real 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.

The technique of *C*-SVM binary classifier[13, 24, 27, 28, 29, 30] has been employed (Fig. 7). The formula created using this technique is the following nonlinear expression with respect to a vector x, which is so-called $y= SVM(x_1, x_2, \ldots, x_m)$ in Figs. 2 and 3:

$$y = \sum_{i=1}^{n}\left[y_i\alpha_i \exp(-\gamma\| \boldsymbol{x} - \boldsymbol{x}_i \|^2)\right] + b \tag{9}$$

where $\boldsymbol{\alpha}$ is the vector of Lagrange multipliers, $\boldsymbol{\alpha}=(\alpha_1, \alpha_2, \ldots, \alpha_n)$, $0\le\alpha_i\le C$ where C is the penalty factor, and the constraint $\sum_{i=1}^{n} y_i\alpha_i=0$; $\exp(-\gamma \| \boldsymbol{x}-\boldsymbol{x}_i \|^2)$ is the RBF (radial basis function) 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 $\boldsymbol{x}_i$. These free $\boldsymbol{x}_i$ are those vectors corresponding to $\alpha_i>0$, on which the final SVM model depends. It is better to take the RBF as a kernel function than to take the linear, polynomial and sigmoid functions under strong nonlinear conditions[30].

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

$$\max_{\boldsymbol{\alpha}}\left\{ \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\cdot\sum_{i,j=1}^{n}\left[\alpha_i\alpha_j y_i y_j \exp(-\gamma\| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2)\right]\right\} \tag{10}$$

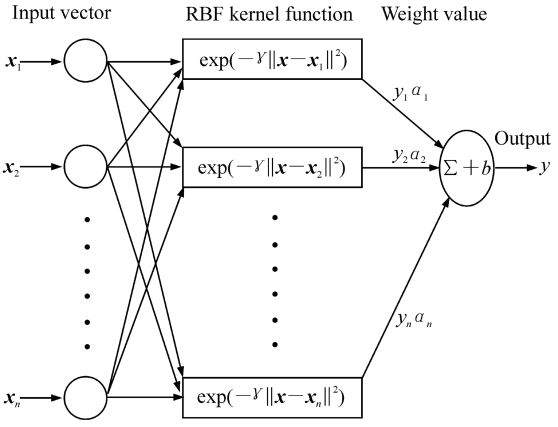

Figure 7
Sketch Map of the Learning Process by *C*-SVM Binary Classifier

1.5 Summary of the Four Classifiers

Each of the above four classifiers follow the same calculating flowchart: a) "Knowledge discovery", i.e., determine a formula, that is, Eq. (4), Eq. (7), Eq. (8), or Eq. (9), using the n learning samples shown by Eq. (1); b) "Knowledge validation", i.e., substitute the n learning samples in the formula to obtain the prediction values $y_1, y_2, \ldots, y_n$ to verify the fit of the classifier; and c) "Knowledge application", i.e., substitute the N prediction samples shown by Eq. (3) in the formula to obtain the prediction values $y_{n+1}, y_{n+2}, \ldots, y_{n+N}$.

The expression $y=y(\boldsymbol{x})$ obtained by MRA, BAYD, BPNN and SVM can be defined as $y=MRA(\boldsymbol{x})$, $y=BAYD(\boldsymbol{x})$, $y=BPNN(\boldsymbol{x})$ and $y=SVM(\boldsymbol{x})$ respectively, where *MRA* is a linear function but *BAYD*, *BPNN* and *SVM* are nonlinear functions. Moreover, *MRA*, *BAYD* and *SVM* are explicit, and can be concretely expressed as mathematical formulas, whereas *BPNN* is implicit.

It is noted for the following two case studies that: since the output y of MRA and BPNN are real-type values whereas the output y of BAYD and SVM are integer-type values, the output y of MRA and BPNN are converted to integer-type values using the Round function for the sake of comparability between the four classifiers.

2. CASE STUDY 1: TRAP QUALITY EVALUATION

Located to the north of the Tarim Basin in western China, the Kuqa Depression covers about 40,000 km^2, stretching from the mountainous southern Tianshan fold belt in the north to the Tabei uplift in the south. It is about 400 km long (E-W) and 50–140 km wide (N-S), wide to the west and narrowing to the east. Over 10 oil and gas fields have been discovered in this depression. It is one of the richest areas of natural gas accumulation in China, of which the large gas field Kela2 has become the major gas supplier in the state project "gas in the west delivered to the east". In the light of the differences of the structural features and migration mechanisms, the Kuqa Depression can be divided into two separate, north and south petroleum systems. The gas-rich Northern Kuqa Depression comprises about half the whole depression and includes the northern monocline belt, the linear Keyi anticline belt, the Qiulitake anticline belt, the Baicheng sag and the Yangxia sag. Since it has experienced stronger tectonic movements, the geological conditions are more complicated than in the Southern Kuqa Depression[11, 31, 32]. Therefore, it is a challenge to study the traps in the Northern Kuqa Depression under such complicated conditions, and it is a new trial to apply MRA, BAYD, BPNN and SVM to the evaluation of trap quality.

The objective of this case study is to conduct the optimal selection of traps using multi-geological factors of oil and gas pool-forming, which has practical value in the stage of rolling exploration.

Using data from the Northern Kuqa Depression[33], 30 traps were selected, of which 27 were taken as learning samples and 3 as prediction samples (Table 1). MRA, BAYD, BPNN and SVM were then applied to trap quality evaluation, using 14 independent variables ($x_1, x_2, \ldots, x_{14}$) and one variable y. In the learning samples, y^* is the input data y assigned by geologists; in the prediction samples y^* (in parenthesis) is also assigned by geologists but it is not used as input data, it is only used for calculating the absolute relative residual R between y and y^* (Table 1). It is worth pointing out that: though geologists had assigned trap quality values to the three prediction samples, these values were judged to be less reliable.

Table 1
Input Data of Trap Quality Evaluation of the Northern Kuqa Depression of the Tarim Basin in Western China

Sample type	Trap No.	x_1	x_2	x_3	x_4 (m)	x_5 (m)	x_6 (km^2)	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14} (Mt)	y*
Learning samples	1	1	1	2	2362	300	58	2	0.45	0.753	0.960	0.935	0.808	0.900	6.6	2
	2	1	2	1.5	3150	350	42	2	0.85	1.000	1.000	0.935	0.921	0.900	210.5	1
	3	1	2	2	3650	350	12	2	0.51	0.975	1.000	0.935	0.763	0.900	8.3	1
	4	1	2	2	2630	150	17	2	0.51	0.818	0.898	0.935	0.763	0.900	1.9	2
	5	1	3	2	5950	750	135	2	0.45	0.895	0.940	0.820	0.808	0.900	171.9	1
	6	1	2	2	3970	300	28	2	0.75	0.950	0.868	0.820	0.763	0.900	5.5	2
	7	1	2	2	4680	300	27	2	0.75	0.828	0.868	0.820	0.808	0.900	12.6	2
	8	1	1	2	1450	700	54	1	0.45	0.778	0.898	0.935	0.751	0.900	7.13	2
	9	1	1	3	1450	1000	74	1	0.45	0.778	0.898	0.935	0.808	0.900	9.8	2
	10	1	2	3	1200	750	23	1	0.45	0.888	0.970	0.840	0.681	0.900	1.4	2
	11	1	1	2	1550	1780	34	2	0.45	0.778	0.860	0.820	0.856	0.900	43.2	2
	12	1	1	2	6700	250	11	2	0.45	0.693	0.930	0.935	0.936	0.900	13.6	1
	13	1	1	2	5500	500	16	2	0.45	0.693	1.000	0.935	0.936	0.900	20.3	1
	14	1	1	2	5500	200	11	2	0.45	0.753	1.000	0.935	0.936	0.900	13.6	1
	15	1	2	1	850	550	50	1	0.45	1.000	0.868	0.820	0.681	0.900	1.82	2
	16	1	2	3	1510	750	57	1	0.45	1.000	1.000	0.840	0.794	0.930	3.48	1
	17	1	1	3	3510	1150	161	2	0.75	0.865	1.000	0.900	0.794	0.930	179.9	1
	18	1	1	3	2700	300	56	1	0.45	0.888	0.970	0.840	0.714	0.900	3.41	2
	19	2	3	1	4220	460	66	2	0.51	0.808	0.882	0.840	0.756	0.930	9.8	2
	20	2	2	3	5600	300	27	2	0.51	0.828	0.882	0.840	0.756	0.930	29.2	2
	21	2	1	3	8580	300	17	2	0.51	0.780	0.898	0.840	0.748	0.930	18.5	2
	22	2	3	1	4940	260	46	2	0.51	0.808	0.798	0.840	0.756	0.930	3.9	2
	23	3	1	1	1855	1800	42	2	0.85	0.865	0.798	0.840	0.909	0.900	4	1
	24	3	3	2	4755	700	83	2	0.85	0.808	1.000	0.850	0.920	0.900	169.4	1
	25	1	2	3	1000	400	82	1	0.45	0.913	0.970	0.885	0.673	0.930	2.61	2
	26	1	1	4	3670	300	133	2	0.45	0.753	0.970	0.885	0.673	0.900	37.9	2
	27	3	3	1	2750	1100	118	2	0.85	0.955	0.898	0.780	0.763	0.900	10.2	2
Prediction samples	28	1	2	3	4450	250	35	2	0.60	0.853	0.898	0.900	0.898	0.900	58.9	(1)
	29	2	3	3	5660	340	56	2	0.51	0.808	0.860	0.850	0.748	0.930	71.8	(2)
	30	2	1	3	5850	180	17	2	0.51	0.753	0.758	0.840	0.728	0.930	19.1	(2)

Note: x_1 = unit structure (1–linear anticline belt, 2–Yangxia sag, 3–Qiulitake anticline belt); x_2 = trap type (1–faulted nose, 2–anticline, 3–faulted anticline); x_3 = petroliferous formation (1–E, 1.5–E+K, 2–K, 3–J, 4–T); x_4 = trap depth; x_5 = trap relief; x_6 = trap closed area; x_7= formation HC identifier (1–oil, 2–gas); x_8 = data reliability (0–1); x_9 = trap coefficient (0–1); x_{10} = source rock coefficient (0–1); x_{11} = reservoir coefficient (0–1); x_{12} = preservation coefficient (0–1); x_{13} = configuration coefficient (0–1); x_{14} = resource quantity (million ton oil-equivalent); y* = trap quality value (1–high,2–low) assigned by geologists.

Using the 27 learning samples (Table 1) and by MRA, BAYD, BPNN, and SVM, the relationship between the predicted trap quality value (y) and the 14 geological factors (x_1, x_2, …, x_{14}) has been determined.

Using MRA, the result $y=MRA(x_1, x_2, \ldots, x_{14})$ obtained is an explicit expression:

$$\begin{aligned} y=&13.766-0.026405x_1-0.038781x_2-0.0016605x_3 \\ &-0.00012343x_4-0.00038344x_5-0.002442x_6+0.045162x_7 \\ &+0.58229x_8-3.3236x_9-2.1313x_{10}-3.3651x_{11}-4.1977x_{12} \\ &-0.67296x_{13}+0.0010516x_{14} \end{aligned} \quad (11)$$

Equation (11) yields a mean square error of 0.14157 and a multiple-correlative coefficient of 0.92652. In addition, the trap quality value (y) is shown to depend on the 14 geological factors in decreasing order: 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.

Using BAYD, the Bayesian discrimination function defined in Eq. (5) obtained is an explicit expression:

$$\left.\begin{aligned} B_1(x) =& \ln(0.37) - 6090.107 - 37.076x_1 + 31.65x_2 \\ &- 11.927x_3 + 0.042x_4 + 0.194x_5 \\ &+ 1.062x_6 + 8.483x_7 - 15.526x_8 \\ &+ 1305.707x_9 + 1175.127x_{10} + 1799.427x_{11} \\ &+ 1870.111x_{12} + 7090.26x_{13} - 1.129x_{14} \\ B_2(x) =& \ln(0.63) - 5746.595 - 37.817x_1 + 30.562x_2 \\ &- 11.974x_3 + 0.038x_4 + 0.184x_5 \\ &+ 0.993x_6 + 9.75x_7 + 0.806x_8 \\ &+ 1212.488x_9 + 1115.348x_{10} + 1705.045x_{11} \\ &+ 1752.377x_{12} + 7077.386x_{13} - 1.099x_{14} \end{aligned}\right\} \quad (12)$$

In addition, the trap quality value (y) is shown to depend on the 14 geological factors in decreasing order: 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. This order is the same as that obtained by the above MRA calculation, which is resulted from the fact that: though MRA is linear while BAYD is nonlinear, the nonlinearity

of the studied problem is quite weak.

The BPNN classifier used consists of 14 input layer nodes, 1 output layer node and 29 hidden layer nodes; the value of the network learning rate for the output layer and the hidden layer is 0.6, the termination of calculation accuracy is 0.0001, and the learning time count is 13120. The result

$$y=BPNN(x_1, x_2, \ldots, x_{14}) \quad (13)$$

obtained is an implicit expression with a global error of 0.00135.

Using *C*-SVM binary classifier, the termination of calculation accuracy is 0.001, C=32 and γ=0.03125, and the cross-validation accuracy is 92.6%. In total, there are 14 free $\boldsymbol{x}_i$. The result

$$y=SVM(x_1, x_2, \ldots, x_{14}) \quad (14)$$

obtained is a nonlinear function that can be concretely expressed as a mathematical formula corresponding to Eq. (9), an explicit expression. It is not, however, reproduced here due to its large size.

Substituting the independent variables determined from the 27 learning samples (Table 1) in Eq. (11), Eq. (12) [and then use Eq. (7)], Eq. (13) and Eq. (14) respectively, the predicted trap quality value of each learning sample for the four classifiers is obtained, thus verifying the predictive accuracy of each classifier. Similarly, the predicted trap quality values of the 3 prediction samples (Table 1) were obtained (Table 2). It can be calculated from Table 2 that for the learning samples, the mean absolute relative residuals between y and y* for the four classifiers are all 0%, and for the prediction samples, 16.67%, 0%, 0%, and 0%, respectively (Table 3).

Table 2
Prediction Results of Trap Quality Evaluation of the Northern Kuqa Depression of the Tarim Basin in Western China

Sample type	Trap No.	y*	Predicted trap quality value							
			MRA		BAYD		BPNN		SVM	
			y	R (%)	y	R (%)	y	R (%)	y	R (%)
Learning samples	1	2	2	0	2	0	2	0	2	0
	2	1	1	0	1	0	1	0	1	0
	3	1	1	0	1	0	1	0	1	0
	4	2	2	0	2	0	2	0	2	0
	5	1	1	0	1	0	1	0	1	0
	6	2	2	0	2	0	2	0	2	0
	7	2	2	0	2	0	2	0	2	0
	8	2	2	0	2	0	2	0	2	0
	9	2	2	0	2	0	2	0	2	0
	10	2	2	0	2	0	2	0	2	0
	11	2	2	0	2	0	2	0	2	0
	12	1	1	0	1	0	1	0	1	0
	13	1	1	0	1	0	1	0	1	0
	14	1	1	0	1	0	1	0	1	0
	15	2	2	0	2	0	2	0	2	0
	16	1	1	0	1	0	1	0	1	0
	17	1	1	0	1	0	1	0	1	0
	18	2	2	0	2	0	2	0	2	0
	19	2	2	0	2	0	2	0	2	0
	20	2	2	0	2	0	2	0	2	0
	21	2	2	0	2	0	2	0	2	0
	22	2	2	0	2	0	2	0	2	0
	23	1	1	0	1	0	1	0	1	0
	24	1	1	0	1	0	1	0	1	0
	25	2	2	0	2	0	2	0	2	0
	26	2	2	0	2	0	2	0	2	0
	27	2	2	0	2	0	2	0	2	0
Prediction samples	28	(1)	1	0	1	0	1	0	1	0
	29	(2)	2	0	2	0	2	0	2	0
	30	(2)	3	50	2	0	2	0	2	0

Note: y* and y are the trap quality value (1–high, 2–low) where y* is assigned by geologists and y is calculated by classifier; R is the absolute relative residual between y and y*.

Table 3
Comparison between the Applications of MRA, BAYD, BPNN, and SVM to Trap Quality Evaluation in the Northern Kuqa Depression of the Tarim Basin in Western China

Classifier	Fitting formula	Mean absolute relative residuals (%)		Dependence of the predicted value (y) on parameters (x_1, x_2, …, x_{14}), in decreasing order	Time consuming on PC (Intel Core 2)	Integrated evaluation
		Learning samples	Prediction samples			
MRA (successive regression analysis)	Linear, explicit	0	16.67	$x_{12}, x_{10}, x_9, x_4, x_5, x_{11}, x_8, x_6, x_{14}, x_1, x_2, x_7, x_{13}, x_3$	<1 s	Average
BAYD (successive Bayesian discrimination)	Nonlinear, explicit	0	0	$x_{12}, x_{10}, x_9, x_4, x_5, x_{11}, x_8, x_6, x_{14}, x_1, x_2, x_7, x_{13}, x_3$	3 s	Excellent
BPNN (back-propagation neural network)	Nonlinear, implicit	0	0	N/A	1 min 20 s	Excellent
SVM (*C*-SVM binary classifier)	Nonlinear, explicit	0	0	N/A	3 s	Excellent

For 27 learning samples and 3 prediction samples (Table 1), the predicted results by all of BAYD, BPNN and SVM are available, which all coincide with trap quality value assigned by geologists (Table 2). BAYD, BPNN and SVM yield values of 1, 2 and 2 for Traps No. 28, 29 and 30 respectively, i.e., the trap quality of No. 28 is high whereas that of both No. 29 and 30 is low. During recent exploration work conducted after the forecast, commercial gas was discovered at Trap No. 28 with 0.7814 tcf (22.127×10^9 m^3) of reserves in place and 0.5470 tcf (15.489×109 m3) of recoverable reserves, but none has yet been discovered at Traps No. 29 and 30. As for MRA, though its mean R is 0% for 27 learning samples, its mean R is 16.67% for 3 prediction samples (Table 3), which is larger than 5% (a standard threshold in general).The fit achieved by MRA is poor, therefore the results can only be used as auxiliary data to show the dependence of the predicted value (y) on parameters (x_1, x_2, …, x_{14}).

3. CASE STUDY 2: OIL IDENTIFICATION

Located on the southwestern margin of the Jianghan Basin in central China, the Xiefengqiao anticlinal structure is a litho-structure complex oil reservoir with low porosity and low permeability lying at depths of 3100–3600 m.

The objective of this case study is to conduct an oil identification of the oil-bearing layers in tight sandstones using conventional well-log data, which has practical value when oil test data is limited.

Using data from the Xiefengqiao anticlinal structure[34], 27 samples were selected, of which 24 were taken as learning samples and 3 as prediction samples (Table 4). MRA, BAYD, BPNN and SVM were then applied to oil identification, using five independent variables: (x_1, x_2, x_3, x_4, x_5) and one variable y. In the learning samples, y^* is the input data y determined by the oil test; in the prediction samples y^* (in parenthesis) is also determined by the oil test but it is not used as input data, it is only used for calculating the absolute relative residual R between y and y^* (Table 4). It is worth pointing out that: since the oil test of the last three samples was judged to be relatively less reliable, these samples are taken as the prediction samples.

Table 4
Input Data of Oil Identification of the Xiefengqiao Anticlinal Structure of the Jianghan Basin in Central China

Sample type	Sample No.	Well No.	Layer No.	x_1 (Ω•m)	x_2 (μs/m)	x_3 (%)	x_4 (%)	x_5 (mD)	y*
Learning samples	1	ES4	5	64	206	6.0	48.6	1.1	2
	2		6	140	208	6.5	41.5	1.3	3
	3		7	63	206	6.0	36.4	1.1	2
	4		8	116	196	3.8	0.7	0.5	3
	5		9	17	267	19.6	44.0	32.4	1
	6		10	49	226	10.6	57.2	5.2	2
	7		11	44	208	6.6	36.1	1.4	2
	8		12	90	208	6.5	29.7	1.3	3
	9		13	69	260	18.1	81.7	16.8	1
	10	ES6	4	49	207	6.2	67.5	0.9	2
	11		5	80	207	6.3	50.9	1.0	3
	12		6	95	218	8.7	77.5	2.2	3
	13		8	164	212	7.5	67.5	1.5	3
	14	ES8	5_1	21	202	5.1	22.2	1.0	2
	15		5_2	56	192	2.9	24.2	0.6	3
	16		5_3	36	198	4.1	28.8	0.8	2
	17		6	128	196	3.4	19.2	0.6	3
	18		11	34	197	3.9	28.4	0.7	2
	19		12_1	10	208	6.4	42.4	1.7	1
	20		12_2	6	226	10.4	45.6	5.8	1
	21		12_3	6	225	10.3	50.4	6.1	1
	22		12_4	10	206	6.0	44.0	1.7	1
	23		13_1	7	224	9.9	44.4	5.2	1
	24		13_2	15	197	3.8	34.2	0.6	1
Prediction samples	25	ES8	13_4	11	201	4.8	39.3	0.8	(1)
	26		13_5	25	197	3.8	16.9	0.6	(2)
	27		7_2	109	199	4.4	17.8	0.8	(3)

Note: x_1 = true resistivity log (RT); x_2 = compensated acoustic log (AC); x_3 = porosity ($\varnothing$); x_4 = oil saturation (S_o); x_5 = permeability (k); y* = oil identification value (1–oil layer, 2–poor oil layer, 3–dry layer) determined by the oil test.

Using the 24 learning samples (Table 4) and by MRA, BAYD, BPNN, and SVM, the relationship between the predicted oil identification value (y) and the five well-logs (x_1, x_2, x_3, x_4, x_5) has been determined.

Using MRA, the result $y=MRA(x_1, x_2, x_3, x_4, x_5)$ obtained is an explicit expression:

$$y=62.641+0.0134x_1-0.3378x_2+1.3781x_3-0.0007x_4+0.0386x_5 \quad (15)$$

Equation (15) yields a mean square error of 0.16649 and a multiple-correlative coefficient of 0.91297. In addition, the oil identification value (y) is shown to depend on the five well-log data in decreasing order: x_1 (RT), x_2 (AC), x_3 (POR), x_5 (k), x_4 (S_o).

Using BAYD, the Bayesian discrimination function defined in Eq. (5) obtained is an explicit expression:

$$\left.\begin{aligned} B_1(x) &= \ln(0.333) - 93842.2 - 3.892x_1 + 1039.76x_2 \\ &\quad - 4468.595x_3 - 0.712x_4 - 73.876x_5 \\ B_2(x) &= \ln(0.333) - 93203.96 - 3.807x_1 + 1036.232x_2 \\ &\quad - 4454.111x_3 - 0.701x_4 - 73.528x_5 \\ B_3(x) &= \ln(0.333) - 92927.46 - 3.674x_1 + 1034.66x_2 \\ &\quad - 4447.812x_3 - 0.729x_4 - 73.279x_5 \end{aligned}\right\} \quad (16)$$

In addition, the oil identification value (y) is shown to depend on the five well-log data in decreasing order: x_1 (RT), x_2 (AC), x_3 (POR), x_5 (k), x_4 (S_o). This order is the same as that obtained by the above MRA calculation, which is resulted from the fact that: though MRA is linear while BAYD is nonlinear, the nonlinearity of the studied problem is quite strong, whereas the ability of nonlinearity of BAYD is low.

The BPNN classifier used consists of 5 input layer nodes, 1 output layer node and 11 hidden layer nodes; the value of the network learning rate for the output layer and the hidden layer is 0.6, the termination of calculation accuracy is 0.0001, and the learning time count is 41467. The result

$$y=BPNN(x_1, x_2, x_3, x_4, x_5) \quad (17)$$

obtained is an implicit expression with a global error of 0.0009111.

Using C-SVM binary classifier, the termination of calculation accuracy is 0.001, C=8192 and γ=0.007813, and the cross-validation accuracy is 91.7%. In total, there are 9 free $\boldsymbol{x}_i$. The result

$$y=SVM(x_1, x_2, x_3, x_4, x_5) \quad (18)$$

obtained is a nonlinear function that can be concretely expressed as a mathematical formula corresponding to Eq. (9), an explicit expression. It is not, however, reproduced here due to its large size.

Substituting the independent variables determined from the 24 learning samples (Table 4) in Eq. (15), Eq. (16) [and then use Eq. (7)], Eq. (17) and Eq. (18) respectively, the predicted oil identification value of each learning sample for the four classifiers is obtained, thus verifying the predictive accuracy of each classifier. Similarly, the predicted oil identification values of the 3 prediction samples (Table 4) were obtained (Table 5). It can be calculated from Table 5 that for the learning samples, the mean absolute relative residuals between y and y^* for the four classifiers are 8.33%, 11.11%, 4.17%, and 0% respectively, and for the prediction samples, 33.33%, 33.33%, 16.67%, and 0% (Table 6).

Table 5
Prediction Results of Oil Identification of the Xiefengqiao Anticlinal Structure of the Jianghan Basin in Central China

Sample type	Trap No.	Well No.	Layer No.	y^*	Predicted oil identification value							
					MRA		BAYD		BPNN		SVM	
					y	R (%)	y	R (%)	y	R (%)	y	R (%)
Learning samples	1	ES4	5	2	2	0	2	0	3	50	2	0
	2		6	3	3	0	3	0	3	0	3	0
	3		7	2	2	0	2	0	3	50	2	0
	4		8	3	3	0	3	0	3	0	3	0
	5		9	1	1	0	1	0	1	0	1	0
	6		10	2	2	0	2	0	2	0	2	0
	7		11	2	2	0	2	0	2	0	2	0
	8		12	3	3	0	3	0	3	0	3	0
	9		13	1	1	0	1	0	1	0	1	0
	10	ES6	4	2	2	0	2	0	2	0	2	0
	11		5	3	2	33.33	3	0	3	0	3	0
	12		6	3	2	33.33	2	33.33	3	0	3	0
	13		8	3	4	33.33	3	0	3	0	3	0
	14	ES8	5_1	2	2	0	2	0	2	0	2	0
	15		5_2	3	3	0	2	33.33	3	0	3	0
	16		5_3	2	2	0	2	0	2	0	2	0
	17		6	3	3	0	3	0	3	0	3	0
	18		11	2	2	0	2	0	2	0	2	0
	19		12_1	1	1	0	1	0	1	0	1	0
	20		12_2	1	1	0	1	0	1	0	1	0
	21		12_3	1	1	0	1	0	1	0	1	0
	22		12_4	1	1	0	2	100	1	0	1	0
	23		13_1	1	1	0	1	0	1	0	1	0
	24		13_2	1	1	0	2	100	1	0	1	0
Prediction samples	25	ES8	13_4	(1)	2	100	2	100	1	0	1	0
	26		13_5	(2)	2	0	2	0	3	50	2	0
	27		7_2	(3)	3	0	3	0	3	0	3	0

Note: y^* and y are the oil identification value (1–oil layer, 2–poor oil layer, 3–dry layer) where y^* is determined by the oil test and y is calculated by classifier; R is the absolute relative residual between y and y^*.

Table 6
Comparison Between the Applications of MRA, BAYD, BPNN, and SVM to Oil Identification of the Xiefengqiao Anticlinal Structure of the Jianghan Basin in Central China

Classifier	Fitting formula	Mean absolute relative residuals (%)		Dependence of the predicted value (y) on parameters (x_1, x_2, x_3, x_4, x_5), in decreasing order	Time consuming on PC (Intel Core 2)	Integrated evaluation
		Learning samples	Prediction samples			
MRA (successive regression analysis)	Linear, explicit	8.33	33.33	x_1, x_2, x_3, x_5, x_4	<1 s	Poor
BAYD (successive Bayesian discrimination)	Nonlinear, explicit	11.11	33.33	x_1, x_2, x_3, x_5, x_4	3 s	Poor
BPNN (back-propagation neural network)	Nonlinear, implicit	4.17	16.67	N/A	1 min 5 s	Average
SVM (*C*-SVM binary classifier)	Nonlinear, explicit	0	0	N/A	3 s	Excellent

For 24 learning samples (Table 4), the predicted results by only SVM are available, which not only all coincide with oil test data (Table 5) but also all with well-log interpretation results[34], though the mean R of BPNN is 4.17% (Table 6) which is less than 5% (a standard threshold in general). Three prediction samples (Table 4), from layers 134, 135 and 72, were mistakenly judged by well-log interpretation to be a dry layer, dry layer and poor oil layer respectively. In fact they are an oil layer, poor oil layer, and dry layer, as validated by oil tests[34]. Table 5 shows that the three results predicted by SVM are all correct, but only two results determined by MRA, BAYD and BPNN are correct with one result being incorrect. This indicates SVM has the ability to correct some erroneous well-log interpretations[34], while MRA, BAYD and BPNN are less accurate. The fit achieved by MRA and BAYD are poor, therefore the results can only be used as auxiliary data to show the dependence of the predicted value (y) on parameters (x_1, x_2, x_3, x_4, x_5).

CONCLUSIONS

From the two case studies of trap quality evaluation and oil identification by using four classifiers (MRA, BAYD, BPNN and SVM), we can draw the following conclusions:

1) Since the nonlinearity of Case 1 is weak, the results of BAYD, BPNN and SVM are same and can have zero residuals, while MRA has unallowable residuals.

2) Since the nonlinearity of Case 2 is strong, the results of only SVM have zero residuals, while BAYD, BPNN and MRA have unallowable residuals.

3) MRA and BAYD can establish the order of dependence between y and each variable (x_1, x_2, …, x_m), which could be used for dimensional reduction in data mining[8] and which cannot be estimated using BPNN and SVM.

4) BPNN is more time-consuming than BAYD and SVM, and MRA runs the fastest.

5) The predictions of the applicable classifiers coincide with real exploration results, and a commercial gas trap was discovered after the forecast in Case 1 and SVM can correct some erroneous well-log interpretations in Case 2.

Therefore, it is proposed that: we can adopt MRA when a problem is linear; adopt BAYD, BPNN, or SVM when a problem is weakly nonlinear; and adopt only SVM when a problem is strongly nonlinear.

REFERENCES

[1] Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA, USA: MIT Press.

[2] Larose, D. T. (2005). *Discovering Knowledge in Data*. New York, USA: John Wiley & Sons, Inc.

[3] Larose, D. T. (2006). *Data Mining Methods and Models*. New York, USA: John Wiley & Sons, Inc.

[4] Hirsh, H. (2008). Data Mining Research: Current Status and Future Opportunities. *Statistical Analysis and Data Mining, 1*(2), 104-107.

[5] Wong, P. M. (2003). A Novel Technique for Modeling Fracture Intensity: A Case Study from the Pinedale Anticline in Wyoming. *AAPG Bulletin, 87*(11), 1717-1727.

[6] Aminzadeh, F. (2005). Applications of AI and Soft Computing for Challenging Problems in the Oil Industry. *Journal of Petroleum Science and Engineering, 47*(1-2), 5-14.

[7] Mohaghegh, S. D. (2005). A New Methodology for the Identification of Best Practices in the Oil and Gas Industry, Using Intelligent Systems. *Journal of Petroleum Science and Engineering, 49*(3-4), 239-260.

[8] Shi, G. R., & Yang, X. S. (2010). Optimization and Data Mining for Fracture Prediction in Geosciences. *Procedia Computer Science, 1*(1), 1353-1360.

[9] Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression*

Analysis by Examples (3rd ed.). New York, USA: John Wiley & Sons, Inc.

[10] Lee, J. H., & Yang, S. H. (2002). Statistical Optimization and Assessment of a Thermal Error Model for CNC Machine Tools. *Int. J. Machine Tools and Manufacture, 42*(1), 147-155.

[11] 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.

[12] Singh, J., Shaik, B., Singh, S., Agrawal, V. K., Khadikar, P. V., Deeb, O., & Supuran, C. T. (2008). Comparative QSAR Study on Para-Substituted Aromatic Sulphonamides as CAII Inhibitors: Information Versus Topological (Distance-Based and Connectivity) Indices. *Chemical Biology and Drug Design, 71*, 244-259.

[13] Shi, G. R. (2009). The Use of Support Vector Machine for Oil and Gas Identification in Low-Porosity and Low-Permeability Reservoirs. *Int. J. Mathematical Modelling and Numerical Optimisation, 1*(1/2), 75-87.

[14] Denison, D. G. T., Holmes, C. C., Mallick, B. K., & Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Chichester, England, UK: John Wiley & Sons, Inc.

[15] Logan, T. P., & Gupta, A. K. (1993). Bayesian Discrimination Using Multiple Observations. *Communications in Statistics-Theory and Methods, 22*(6), 1735-1754.

[16] Brown, P. J., Kenward, M. G., & Bassett, E. E. (2001). Bayesian Discrimination with Longitudinal Data. *Biostatistics, 2*(4), 417-432.

[17] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning Internal Representations by Error Propagation*. In Parallel Distributed Processing, D. E. Rumelhart, & J. L. McClelland (Eds.), Volume 1, pp. 317-362. Cambridge, MA, USA: MIT Press.

[18] Hecht-Nielsen, R. (1989). *Theory of the Backpropagation Neural Network*. In Proceedings of the Int. Joint Conf. on Neural Networks (pp. 593-605), Washington.

[19] Güler,i., & Übeyli, E. D. (2003). Detection of Ophthalmic Artery Stenosis by Least-Mean Squares Backpropagation Neural Network. *Computers in Biology and Medicine, 33*(4), 333-343.

[20] Altiparmak, F., Dengiz, B., & Bulgak, A. A. (2007). Buffer Allocation and Performance Modeling in Asynchronous Assembly System Operations: An Artificial Neural Network Metamodeling Approach. *Applied Soft Computing, 7*(3), 946-956.

[21] Tabach , E. E., Lancelot, L., Shahrour, I., & Najjar, Y. (2007). Use of Artificial Neural Network Simulation Metamodelling to Assess Groundwater Contamination in a Road Project. *Mathematical and Computer Modelling, 45*(7-8), 766-776.

[22] Choi, B., Lee, J. H., & Kim, D. H. (2008). Solving Local Minima Problem with Large Number of Hidden Nodes on Two-Layered Feed-Forward Artificial Neural Networks. *Neurocomputing, 71*(16-18), 3640-3643.

[23] Bennett, K. P., & Mangasarian, O. L. (1992). Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. *Optimization Methods and Software, 1*(1), 23-34.

[24] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York, USA: Springer-Verlag.

[25] Blum, A. L., & Langley, P. (1997). Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence, 97*(1-2), 245-271.

[26] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines, *Machine Learning, 46*(1-3), 389-422.

[27] Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge, UK: Cambridge University Press.

[28] Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New Support Vector Algorithms. *Neural Computation, 12*(5), 1207-1245.

[29] Shi, G. R. (2008). Superiorities of Support Vector Machine in Fracture Prediction and Gassiness Evaluation. *Petroleum Exploration and Development. 35*(5), 588-594.

[30] Chang, C. C., & Lin, C. J. (2011). *LIBSVM: a library for support vector machines, Version 3.1*. Retrived from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[31] Zhou, X. X., Zhang, G. Y., Li, H. H., Wang, H. J., & Jia, J. H. (2002). *Factors of Pool-forming of the Kuqa Petroleum System in the Tarim Basin*. Beijing, China: Petroleum Industry Press. (in Chinese)

[32] Shi, G. R., Ma J. S., Yang X. S., Chang J. H., & Wan, J. (2011). Finite Volume Method for Solving a Modified 3-D 3-Phase Black-Oil Hydrocarbon Secondary Migration Model, and Its Application to the Kuqa Depression of the Tarim Basin in Western China. *Advances in Petroleum Exploration and Development, 2*(1), 1-12.

[33] Shi, G. R., Zhang, G. Y., & Shi, X. F. (2002). Application of Artificial Neural Network and Multiple Regression Analysis to Optimization of Exploration Prospects. *Acta Petrolei Sinica, 23*(5), 19-22. (in Chinese with English abstract)

[34] Yang, J. X. (2002). Identification of Oil Horizons by Artificial Neural Networks in Xiefengqiao Structure. *Oil & Gas Geology, 23*(1), 76-80. (in Chinese with English abstract)

