CSCanada
Advances in Petroleum Exploration and Development
Vol. 7, No. 2, 2014, pp. 1-8
DOI:10.3968/5122

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

A Big Data Mining in Petroleum Exploration and Development

SHI Guangren[a],*; ZHU Yixiang[a]; MI Shiyun[a]; MA Jinshan[a]; WAN Jun[a]

[a]Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, China.
*Corresponding author.

Supported by the Research Institute of Petroleum Exploration and Development (RIPED) and PetroChina.

Received 12 May 2014; accepted 24 June 2014
Published onine 28 June 2014

Abstract

We take a well log in petroleum exploration and development as an example of the big data mining, and adopt three regression and two classification algorithms: the multiple regression analysis (MRA), the error back-propagation neural network (BPNN), the regression of support vector machine (*R*-SVM), the classification of support vector machine (*C*-SVM), and the Bayesian successive discrimination (BAYSD). It is well known that MRA, BPNN and *R*-SVM are regression algorithms while *C*-SVM and BAYSD are classification algorithms, and only MRA is linear algorithm whereas the other four algorithms are nonlinear algorithms. From this case study, we can draw the following five major conclusions: a) Since *C*-SVM is the best classifier, it is employed as a data cleaning tool. b) Since MRA is a linear algorithm, its total mean absolute relative residual $\bar{R}^*$(%) can express the nonlinearity of studied problem. For this case study, $\bar{R}^*$(%)=52.14 showing the nonlinearity of the studied problem is strong. c) Since both MRA and BAYD can establish the order of dependence between a dependent variable and independent variables, each of MRA and BAYD could serve as a pioneering dimension-reduction tool in data mining. In the case study, since the nonlinearity of the studied problem is strong, the nonlinear algorithm BAYSD can serve as a pioneering dimension-reduction tool, but the linear algorithm MRA cannot. d) Since the nonlinearity of the case study is strong, BPNN and *R*-SVM are not applicable though they are nonlinear algorithms, whereas other two nonlinear algorithms *C*-SVM and BAYSD are applicable, indicating the nonlinear ability of *C*-SVM and BAYSD is higher than that of BPNN and *R*-SVM. e) Comparing the two applicable algorithms *C*-SVM and BAYSD for this case study, it is seen that $\bar{R}^*$(%) of *C*-SVM is less than that of BAYSD; BAYSD can serve as a pioneering dimension-reduction tool, but *C*-SVM cannot; it is easy to code the BAYSD program whereas it is very complicated to code the *C*-SVM program, so BAYSD is a good software for this case study when *C*-SVM is not available.

Key words: Big data mining; Well log; Data cleaning; Dimension-reduction; Regression; Classification

Shi, G. R., Zhu, Y. X., Mi, S. Y., Ma, J. S., & Wan, J. (2014). A big data mining in petroleum exploration and development. *Advances in Petroleum Exploration and Development*, *7*(2), 1-8. Available from: URL: http://www.cscanada.net/index.php/aped/article/view/5122
DOI: http://dx.doi.org/10.3968/5122

INTRODUCTION

In the recent years, the big data mining (BDM) has seen enormous success, in some fields of business and sciences, but the BDM application to petroleum exploration and development (PED) is still in initial stage. This is because the PED is different from the other fields, with miscellaneous data types, huge quantity, different measuring precision, and lots of uncertainties to data mining results. In the PED, the seismic, remote sensing and well log data are potential applications of the BDM. This paper presents a BDM in well log data as an example.

The study of the lithology of volcanic rocks started very early, but most of studies are based on petrochemistry and geochemistry[1, 2]. From the development of volcanic rocks and its controlling on oil/gas reservoirs, we divide the volcanic rocks in the Nioudong Oilfield of the Malang Sag of the Santanghu Basin in NW Chin into 9 types (Table 1), and select 15 lithology-sensitive well logs for the BDM.

Table 1
The Lithologic Code of Volcanic Rock in the Nioudong Oilfield

Lithology code	Lithology	Physical property	Evidences of oil and gas
1	Stoma-amygdaloidal basalt	Best	Best
2	Alterative stoma-amygdaloidal basalt	Good	Good
3	Embrittled basalt	Fracture developed	Good
4	Embrittled tuffaceous breccia		Detected
5	Tuffaceous breccia		A little
6	Stoma-amygdaloidal andesite		A little bit
7	Dense basalt		Not detected
8	Dense andesite		Not detected
9	Tuff		Not detected

1. COMMON USED REGRESSION AND CLASSIFICATION ALGORITHMS

Here introduces in three regression and two classification algorithms: the multiple regression analysis (MRA), the error back-propagation neural network (BPNN), the regression of support vector machine (*R*-SVM), the classification of support vector machine (*C*-SVM), and the Bayesian successive discrimination (BAYSD). These five algorithms use the same known parameters, and also share the same unknown that is predicted. The only difference between them is the method and calculation results. It is well known that MRA, BPNN and *R*-SVM are regression algorithms with real number results while *C*-SVM and BAYSD are classification algorithms with integer number results. Since the case study below is a classification problem, we approximately regard the three regression algorithms as classification algorithms, the results y of the regression algorithms are converted from real number to integer number by using round rule, certainly, it is possible that some y after the conversion are not equal to any observed values y^* in all learning samples.

Assume that there are n learning samples, each associated with m + 1 numbers ($x_1, x_2, \ldots, x_m, y^*_i$) and a set of observed values ($x_{1i}, x_{2i}, \ldots, x_{mi}, y^*_i$), with $i = 1, 2, \ldots, n$ 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:

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im}, y^*_i) \quad (i = 1, 2, \ldots, n) \tag{1}$$

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

Equation (1) is the expression of learning samples.

Let $\boldsymbol{x}_0$ be the general form of a vector of ($x_{i1}, x_{i2}, \ldots, x_{im}$). The principles of MRA, BPNN and BAYSD are the same, i.e. try to construct an expression, $y = y(\boldsymbol{x}_0)$, such that Equation (2) is minimized. Certainly, these three different algorithms use different approaches and result in differing accuracy of calculation results.

$$\sum_{i=1}^{n}\left[y\left(\boldsymbol{x}_{0i}\right)-y_i^*\right]^2 \tag{2}$$

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

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

This $y = y(\boldsymbol{x}_0)$ is called the fitting formula obtained in the learning process. The fitting formulas of different algorithms 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 into the fitting formula to get prediction values ($y_1, y_2, \ldots, y_n$), respectively, so as to verify the fitness of a algorithm; and the 3rd step is the prediction process, substituting k prediction samples expressed with Equation (3) into the fitting formula to get prediction values ($y_{n+1}, y_{n+2}, \ldots, y_{n+k}$), respectively.

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

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

Equation (3) is the expression of prediction samples.

1.1 Linear and Nonlinear Algorithms

In the aforementioned five algorithms, only MRA is a linear algorithm whereas the other four are nonlinear algorithms, this is due to the fact that MRA constructs a linear function whereas the other four construct nonlinear functions, respectively. However, MRA can serve as an auxiliary tool, e.g. a pioneering dimension-reduction tool, cooperating with major tools (BPNN, *R*-SVM, *C*-SVM, and BAYSD) for data mining. Besides MRA, BAYSD also can play an important role as a pioneering dimension-reduction tool, because these two algorithms all can give the dependence of the predicted value (y) on parameters ($x_1, x_2, \ldots, x_m$), in decreasing order. However, because MRA belongs to data analysis in linear correlation whereas BAYSD is in nonlinear correlation, in the applications with very strong nonlinearity the ability of dimension-reduction of BAYSD is higher than that of MRA.

1.2 Error Analysis of Calculation Results

To express the calculation accuracy of the prediction variable y for learning and prediction samples when the aforementioned five algorisms are used, the absolute relative residual $R(\%)$, the mean absolute relative residual $\bar{R}(\%)$ and the total mean absolute relative residual $\bar{R}^*(\%)$ are adopted.

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

$$R(\%)_i \quad \left|(y \quad y_i^*) \quad y_i^*\right| \times 100 \tag{4}$$

where y_i is the calculation result of the dependent variable in the i^{th} learning sample; and the other symbols have been defined in Equations (1) and (3).

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

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

$$\bar{R}(\%) = \sum_{i=1}^{N_s} R(\%)_i \;/\; N_s \tag{5}$$

where $N_s = n$ for learning samples while $N_s = k$ for prediction samples; and the other symbols have been defined in Equations (1) and (3).

For learning samples, $R(\%)$ and $\bar{R}(\%)$ are called the fitting residual to express the fitness of learning process, and here $\bar{R}(\%)$ is designated as $\bar{R}_1(\%)$; and for prediction samples, $R(\%)$ and $\bar{R}(\%)$ are called the prediction residual to express the accuracy of prediction process, and here $\bar{R}(\%)$ is designated as $\bar{R}_2(\%)$.

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

$$\bar{R}^*(\%) = [\bar{R}_1(\%) + \bar{R}_2(\%)]/2 \tag{6}$$

when there are no prediction samples, $\bar{R}^*(\%) = \bar{R}_1(\%)$.

1.3 Nonlinearity and Solution Accuracy of Studied Problem

Since MRA is a linear algorithm, its $\bar{R}^*(\%)$ for a studied problem expresses the nonlinearity of $y = y(\boldsymbol{x})$ to be solved, i.e. the nonlinearity of the studied problem.

Whether linear algorithm (MRA) or nonlinear algorithms (BPNN, *R*-SVM, *C*-SVM, and BAYSD), their $\bar{R}^*(\%)$ of a studied problem expresses the accuracy of $y = y(x)$ obtained by each algorithm, i.e. solution accuracy of the studied problem solved by each algorithm.

$y = y(\boldsymbol{x})$ created by BPNN is an implicit expression, which cannot be expressed as a usual mathematical formula; whereas that of the other four algorithms are explicit expressions, which are expressed as a usual mathematical formula.

2. CASE STUDY: LITHOLOGIC DIVISION OF VOLCANIC ROCK

In the Nioudong Oilfield, the divided 9 types of volcanic rocks are as shown as Table 1; the selected 15 well logs are: acoustictime (AC), bulk density (DEN), photoelectric absorption cross section index (PE), natural gamma ray (GR), compensated neutron (CNL), borehole diameter (CAL), shallow resistivity (RMLL), middle resistivity (RS), deep resistivity (RD), fracture development index (FRCT), density porosity (PORD), air voids porosity index (VUGP), acoustic porosity (POS2), low angle fracture development index (FR_H), and permeability (KALL).

2.1 Input Data

We take three sections for study: 1,410 - 1,580 m of Well N9-10, 1,410 - 1,520 m of Well N9-91, and 1,440 - 1,560 m of Well N8-10 (Figures 1, 2 and 3).

Sampling in 0.125 m interval, these three sections have 1,361, 881 and 961 samples, respectively (Table 2). Each sample contains 15 independent variables (x_1, x_2, …, x_{15}), and one variable (y). That is, x_1 is AC, μs/m; x_2 is DEN, g/cm^3; x_3 is PE, b/e; x_4 is GR, API; x_5 is CNL, %; x_6 is CAL, cm; x_7 is RMLL, Ω·m; x_8 is RS, Ω·m; x_9 is RD, Ω·m; x_{10} is FRCT, 0 - 1; x_{11} is PORD, %; x_{12} is VUGP, %; x_{13} is POS2, %; x_{14} is FR_H; x_{15} is KALL, mD. Moreover, y is prediction value of the lithology of volcanic rocks, expressing in 1, 2, 3, 4, 5, 6, 7, 8, and 9 (Tables 1 and 2). It muse be noticed that a) the lithology code in Figures 1, 2 and 3 is original, but b) in Table 2 is after data cleaning which will be introduced below.

Table 2
The Number of Samples for the Lithologic Code of Volcanic Rocks in Three Sections of Well N9-10, Well N9-91 and Well N8-10

Well	Section (m)	Total number of samples	Total number of samples for each lithologic code								
			Code 1	Code 2	Code 3	Code 4	Code 5	Code 6	Code 7	Code 8	Code 9
N9-10	1,410 - 1,580	1,361	367	76	220	264	216	10	89	4	115
N9-91	1,410 - 1,520	881	269	53	37	81	189	41	26	96	89
N8-10	1,440 - 1,560	961	334	0	106	25	79	222	14	131	50

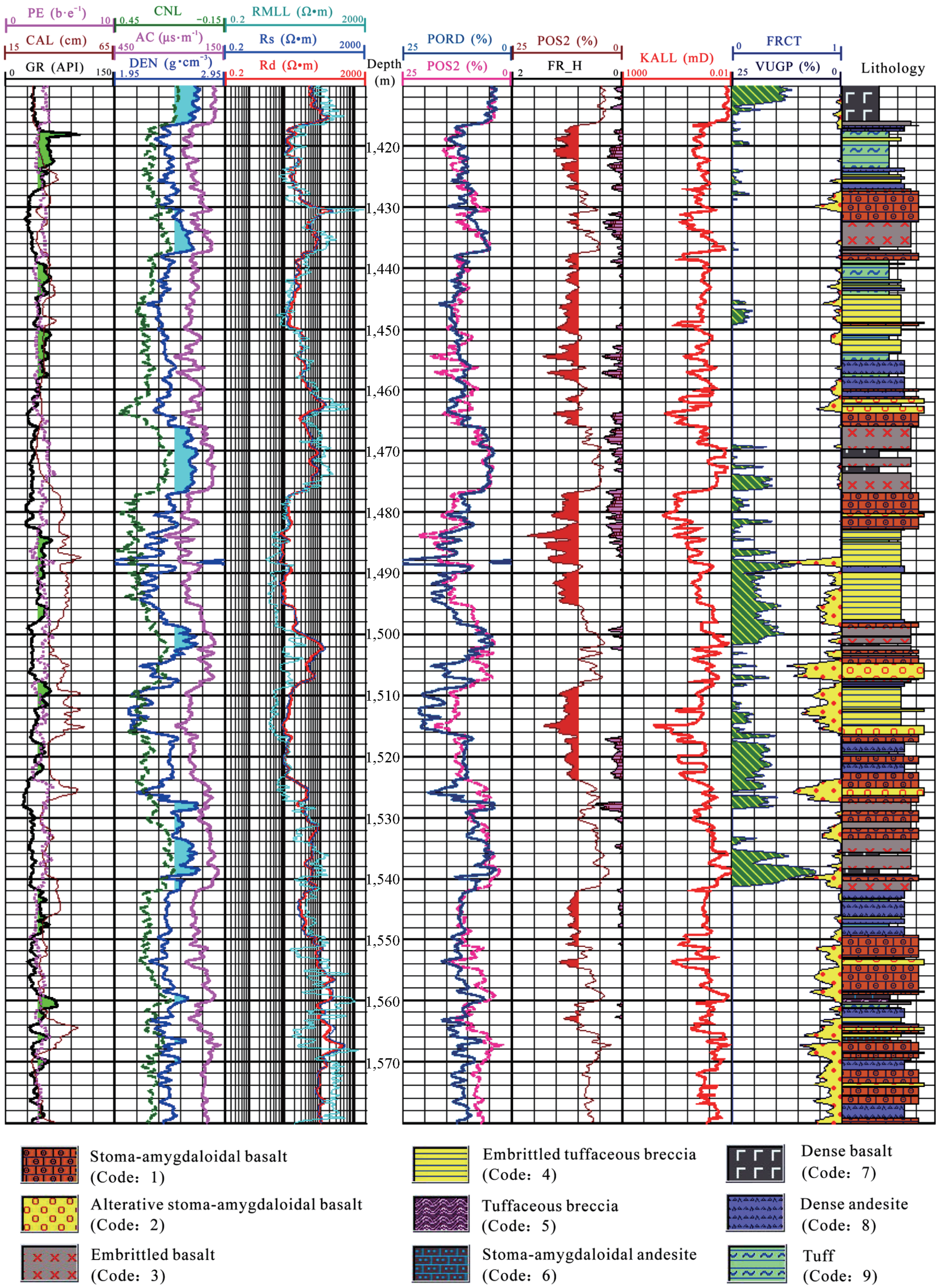

Figure 1
The Synthetic Map of Well Log Interpretation for the Volcanic Rock Reservoirs of Well N9-10 (Modified From [3])

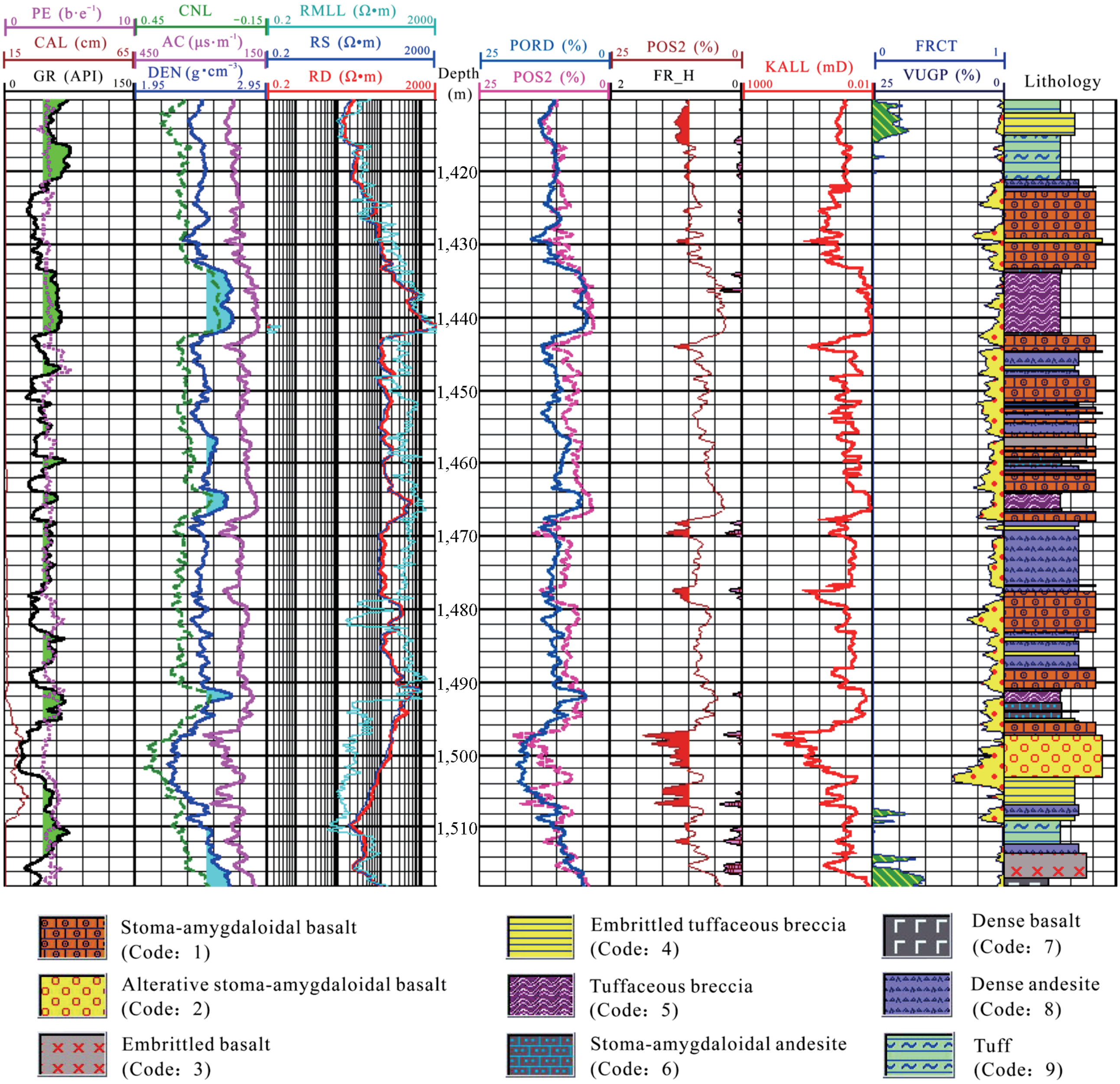

Figure 2
The Synthetic Map of Well Log Interpretation for the Volcanic Rock Reservoirs of Well N9-91 (Modified From [3])

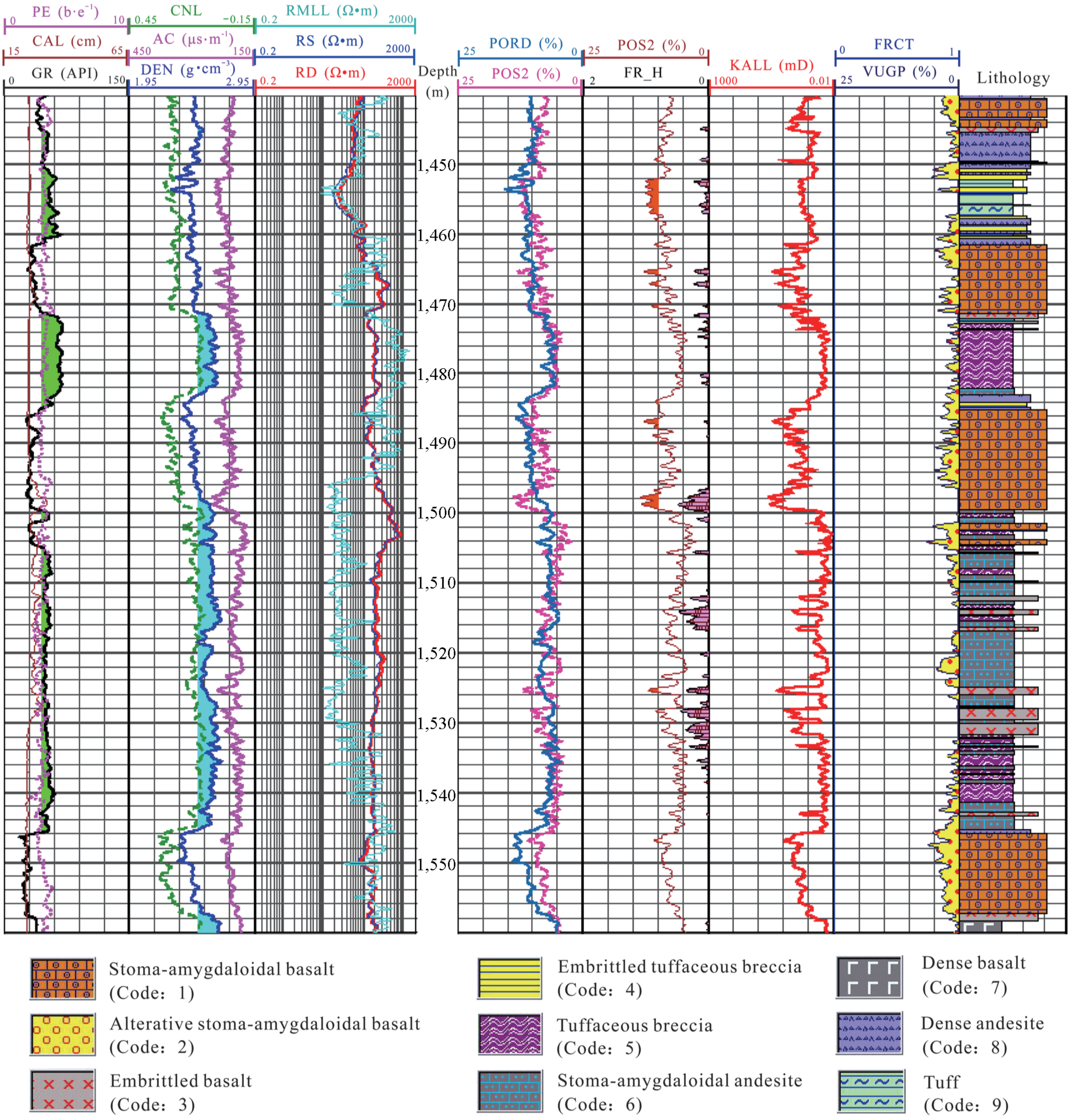

Figure 3
The Synthetic Map of Well Log Interpretation for the Volcanic Rock Reservoirs of Well N8-10 (Modified From [3])

In each of the three sections (Table 2), the numbers of samples for some lithologic codes are less than 50, especially there is no Code 2 in N8-10 section, and so each section cannot be used as the leaning section. Therefore, we combined N9-10 and N9-91 sections as the leaning section, in which there are 2,242 leaning samples and the number of samples for each lithologic code is all more than 50. And we take N8-10 section as the prediction section, in which there are 961 prediction samples, and the lithologic code in each sample is not input data, but is used for calculating *R*(%).

2.2 Data Cleaning

Data cleaning is one of approaches in data preprocessing. Realistic data are often noisy, imperfect and inconsistent. The main job for data cleaning is to fill up the missed data value, make noisy data smooth, identify or eliminate abnormal value as well as solve inconsistent problems. Generally, the process for data cleaning is to handle the missed data first, followed by processing noisy data, and to solve the inconsistent data at last. In this case study, we have employed *C*-SVM, the best classifier[4], to do the learning process on the combined N9-10 and N9-91 sections, and found that $R(\%)$ of 27 samples are not zero since the lithologic code are not given accurately. By correcting them, we have run C-SVM again and got $\bar{R}_1(\%) = 0$ (Table 3).

Table 3
Comparison Between the Applications of MRA, BPNN, *R*-SVM, *C*-SVM and BAYSD

Algorithm	Fitting formula	Mean absolute relative residual			Dependence of the predicted value (y) on parameters (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}), in decreasing order	Time consuming on PC (Intel Core 2)	Solution accuracy
		$\bar{R}_1(\%)$	$\bar{R}_2(\%)$	$\bar{R}^*(\%)$			
MRA	Linear, explicit	51.84	52.44	52.14	x_4, x_{11}, x_{10}, x_{13}, x_5, x_7, x_6, x_{14}, x_{12}, x_3, x_2, x_{15}, x_9, x_8, x_1	5 s	Low
BPNN	Nonlinear, implicit	48.66	46.31	47.49	N/A	20 min	Low
R-SVM	Nonlinear, explicit	39.74	52.00	45.87	N/A	1 min	Low
C-SVM	Nonlinear, explicit	0	6.30	3.15	N/A	4 min	Very high
BAYSD	Nonlinear, explicit	13.89	19.73	16.81	x_{11}, x_4, x_8, x_5, x_{10}, x_{15}, x_1	4 min	High

2.3 Learning Process

Using the 2,242 learning samples (Figures 1 and 2) and by the learning process of MRA, BPNN, R-SVM, C-SVM and BAYSD, to construct an expression, $y = y(x_1, x_2, \ldots, x_{15})$, respectively.

Using MRA[4], the result is an explicit linear function:

$$\begin{aligned} y = {} & 327.7 - 0.0705x_1 - 111.4x_2 - 0.1901x_3 + 0.1676x_4 \\ & - 8.06x_5 - 0.008084x_6 - 0.000353x_7 - 0.002749x_8 \\ & + 0.002731x_9 + 2.736x_{10} - 3.834x_{11} - 0.3925x_{12} \\ & + 0.2911x_{13} + 5.62x_{14} + 0.01414x_{15} \qquad (7) \end{aligned}$$

Equation (7) yields a residual variance of 0.346 and a multiple correlation coefficient of 0.8087, and $\bar{R}^*(\%) = 52.14$ (Table 3) showing the nonlinearity of the studied problem is strong. From the regression process, the lithology of volcanic rocks (y) is shown to depend on the 15 parameters in decreasing order: x_4 (GR), x_{11} (PORD), x_{10} (FRCT), x_{13} (POS2), x_5 (CNL), x_7 (RMLL), x_6 (CAL), x_{14} (FR_H), x_{12} (VUGP), x_3 (PE), x_2 (DEN), x_{15} (KALL), x_9 (RD), x_8 (RS), and x_1 (AC).

The BPNN[4] used consists of 15 input layer nodes, 1 output layer node and 31 hidden layer nodes. The network learning rate of output layer α = 0.6, the network learning rate of hidden layer β = 0.6, and termination of calculation accuracy TCA = 10^{-4}. Thus, the calculated optimal learning time count $t_{\text{opt}} = 5{,}883$, and the result is an implicit nonlinear function:

$$y = \text{BPNN}(x_1, x_2, \ldots, x_{15}) \qquad (8)$$

with the root mean-square error $RMSE(\%) = 0.5863 \times 10^{-1}$.

Equation (8) cannot be expressed as a usual mathematical formula, and so is an implicit expression.

In order to have the comparability of results between R-SVM and C-SVM, RBF is taken as a kernel function, and TCA is fixed to 10^{-3}. Moreover, the insensitive function e in R-SVM is fixed to 0.1.

Using R-SVM[4, 5], the result is an explicit nonlinear function:

$$y = R\text{-}SVM(x_1, x_2, \ldots, x_{15}) \qquad (9)$$

with $C = 1$, $\gamma = 0.1$, and 2,060 free vectors $\boldsymbol{x}_i$.

Using C-SVM[4, 5], the result is an explicit nonlinear function:

$$y = C\text{-}SVM(x_1, x_2, \ldots, x_{15}) \qquad (10)$$

with $C = 2{,}048$, $\gamma = 0.125$, 284 free vectors $\boldsymbol{x}_i$, and the cross validation accuracy CVA = 97.3238%.

Equations (9) and (10) can be expressed as usual mathematical formulas, but are not concretely written out due to their large size.

Using BAYSD[4], the result is a discrimination function:

$$\begin{aligned} B_1 = {} & \ln(0.284) - 167.068 + 1.071\mathrm{x}_1 + 0.0\mathrm{x}_2 + 0.0\mathrm{x}_3 \\ & + 1.83\mathrm{x}_4 - 18.455\mathrm{x}_5 + 0.0\mathrm{x}_6 + 0.0\mathrm{x}_7 + 0.056\mathrm{x}_8 \\ & + 0.0\mathrm{x}_9 + 31.604\mathrm{x}_{10} + 2.78\mathrm{x}_{11} + 0.0\mathrm{x}_{12} + 0.0\mathrm{x}_{13} \\ & + 0.0\mathrm{x}_{14} - 3.946\mathrm{x}_{15} \\ B_2 = {} & \ln(0.058) - 181.807 + 0.991\mathrm{x}_1 + 0.0\mathrm{x}_2 + 0.0\mathrm{x}_3 \\ & + 1.786\mathrm{x}_4 - 15.325\mathrm{x}_5 + 0.0\mathrm{x}_6 + 0.0\mathrm{x}_7 + 0.057\mathrm{x}_8 \\ & + 0.0\mathrm{x}_9 + 28.846\mathrm{x}_{10} + 4.518\mathrm{x}_{11} + 0.0\mathrm{x}_{12} + 0.0\mathrm{x}_{13} \\ & + 0.0\mathrm{x}_{14} - 3.125\mathrm{x}_{15} \\ B_3 = {} & \ln(0.115) - 148.414 + 1.094\mathrm{x}_1 + 0.0\mathrm{x}_2 + 0.0\mathrm{x}_3 \\ & + 1.767\mathrm{x}_4 - 51.178\mathrm{x}_5 + 0.0\mathrm{x}_6 + 0.0\mathrm{x}_7 + 0.049\mathrm{x}_8 \\ & + 0.0\mathrm{x}_9 + 38.693\mathrm{x}_{10} + 1.191\mathrm{x}_{11} + 0.0\mathrm{x}_{12} + 0.0\mathrm{x}_{13} \\ & + 0.0\mathrm{x}_{14} - 3.946\mathrm{x}_{15} \\ B_4 = {} & \ln(0.154) - 244.031 + 1.162\mathrm{x}_1 + 0.0\mathrm{x}_2 + 0.0\mathrm{x}_3 \\ & + 2.53\mathrm{x}_4 + 6.7\mathrm{x}_5 + 0.0\mathrm{x}_6 + 0.0\mathrm{x}_7 + 0.056\mathrm{x}_8 \\ & + 0.0\mathrm{x}_9 + 38.709\mathrm{x}_{10} + 4.177\mathrm{x}_{11} + 0.0\mathrm{x}_{12} + 0.0\mathrm{x}_{13} \\ & + 0.0\mathrm{x}_{14} - 4.571\mathrm{x}_{15} \\ B_5 = {} & \ln(0.181) - 203.932 + 1.12\mathrm{x}_1 + 0.0\mathrm{x}_2 + 0.0\mathrm{x}_3 \\ & + 2.382\mathrm{x}_4 - 5.217\mathrm{x}_5 + 0.0\mathrm{x}_6 + 0.0\mathrm{x}_7 + 0.054\mathrm{x}_8 \\ & + 0.0\mathrm{x}_9 + 33.715\mathrm{x}_{10} + 2.699\mathrm{x}_{11} + 0.0\mathrm{x}_{12} + 0.0\mathrm{x}_{13} \\ & + 0.0\mathrm{x}_{14} - 4.311\mathrm{x}_{15} \\ B_6 = {} & \ln(0.023) - 193.691 + 1.085\mathrm{x}_1 + 0.0\mathrm{x}_2 + 0.0\mathrm{x}_3 \\ & + 2.487\mathrm{x}_4 - 45.016\mathrm{x}_5 + 0.0\mathrm{x}_6 + 0.0\mathrm{x}_7 + 0.065\mathrm{x}_8 \\ & + 0.0\mathrm{x}_9 + 33.858\mathrm{x}_{10} + 2.495\mathrm{x}_{11} + 0.0\mathrm{x}_{12} + 0.0\mathrm{x}_{13} \\ & + 0.0\mathrm{x}_{14} - 3.941\mathrm{x}_{15} \\ B_7 = {} & \ln(0.051) - 141.563 + 1.105\mathrm{x}_1 + 0.0\mathrm{x}_2 + 0.0\mathrm{x}_3 \\ & + 1.728\mathrm{x}_4 - 97.476\mathrm{x}_5 + 0.0\mathrm{x}_6 + 0.0\mathrm{x}_7 + 0.049\mathrm{x}_8 \\ & + 0.0\mathrm{x}_9 + 51.573\mathrm{x}_{10} + 0.67\mathrm{x}_{11} + 0.0\mathrm{x}_{12} + 0.0\mathrm{x}_{13} \\ & + 0.0\mathrm{x}_{14} - 3.776\mathrm{x}_{15} \end{aligned}$$

$$\begin{aligned}B_8 &= \ln(0.045) - 218.589 + 1.21x_1 + 0.0x_2 + 0.0x_3 \\ &+ 2.521x_4 - 98.965x_5 + 0.0x_6 + 0.0x_7 + 0.093x_8 \\ &+ 0.0x_9 + 41.555x_{10} + 1.462x_{11} + 0.0x_{12} + 0.0x_{13} \\ &+ 0.0x_{14} - 4.038x_{15} \\ B_9 &= \ln(0.091) - 252.281 + 1.207x_1 + 0.0x_2 + 0.0x_3 \\ &+ 2.873x_4 - 12.054x_5 + 0.0x_6 + 0.0x_7 + 0.052x_8 \\ &+ 0.0x_9 + 35.768x_{10} + 2.856x_{11} + 0.0x_{12} + 0.0x_{13} \\ &+ 0.0x_{14} - 4.575x_{15}\end{aligned} \tag{11}$$

Then $y = l_b$ if $B_{l_b} = \max\limits_{1 \le l \le 9} \{B_l\}$. Thus, it is believed that an explicit nonlinear function is obtained:

$$y = BAYSD(x_1, x_2, \ldots, x_{15}) \tag{12}$$

From the discriminate process of BAYSD, the lithology of volcanic rocks (y) is shown to depend on the 7 parameters in decreasing order: x_{11} (PORD), x_4 (GR), x_8 (RS), x_5 (CNL), x_{10} (FRCT), x_{15} (KALL), and x_1 (AC). As for the other 8 parameters, they are not introduced, and thus their corresponding coefficients are all zero. This order is quite different from that by MRA (Table 3), because of the fact that MRA is a linear algorithm while BAYSD is a nonlinear algorithm, but the nonlinearity of the studied problem.

Substituting independent variables ($x_1, x_2, \ldots, x_{15}$) of 2,242 learning samples (Figures 1 and 2) into Equations (7), (8), (9), (10) and (12), respectively, the variable y of each learning sample for each algorithm is obtained, and $\bar{R}_1$(%) can be calculated out (Table 3) to show the fitness of each algorithm.

2.4 Prediction Process

Substituting independent variables ($x_1, x_2, \ldots, x_{15}$) of 961 prediction samples (Figure 3) into Equations (7), (8), (9), (10) and (12), respectively, the variable y of each prediction sample for each algorithm is obtained, and $\bar{R}_2$(%) can be calculated out (Table 3) to show the prediction accuracy of each algorithm.

From $\bar{R}_1$(%) and $\bar{R}_2$(%), we get $\bar{R}^*$(%) to express the solution accuracy of each algorithm (Table 3).

2.5 Dimension-Reduction

Dimension-reduction refers to the reduction of the number of independent variables. From Table 3, we know each of MRA and BAYSD could serve as a pioneering dimension-reduction tool in data mining, since they can give the dependence of y on independent variables ($x_1, x_2, \ldots, x_{15}$) in decreasing order. For MRA, x_1 (AC) is the minimum dependence of y (Table 3), so we tried to delete x_1 and run C-SVM, the results shows $\bar{R}_1$(%) = 0.036 and $\bar{R}_2$(%) = 8.13, but the results without this deletion are $\bar{R}_1$(%) = 0 and $\bar{R}_2$(%) = 6.30 (Table 3), which indicates this dimension-reduction is failed. For BAYSD, however, though it runs in the condition without x_{13} (POS2), x_7 (RMLL), x_6 (CAL), x_{14} (FR_H), x_{12} (VUGP), x_3 (PE), x_2 (DEN) and x_9 (RD), its solution accuracy is high (Table 3), which shows the dimension of this studied problem can be reduced from 16-D to 8-D. Why is the dimension-reduction of BAYSD successful but that of MRA failed? The reason is that the nonlinearity of the studied problem is strong due to $\bar{R}^*$(%) of MRA is 52.14, and BAYSD is a nonlinear algorithm while MRA is a linear algorithm.

CONCLUSIONS

From the case study of lithologic division of volcanic rock by using three regression algorithms (MRA, BPNN, R-SVM) and two classification algorithms (C-SVM, BAYSD), in which only MRA is linear algorithm whereas the other four algorithms are nonlinear algorithms, we can draw the following five major conclusions:

(a) Since C-SVM is the best classifier, it is employed as a data cleaning tool.

(b) Since MRA is a linear algorithm, its total mean absolute relative residual $\bar{R}^*$(%) can express the nonlinearity of studied problem. For this case study, $\bar{R}^*$(%)=52.14 showing the nonlinearity of the studied problem is strong.

(c) Since both MRA and BAYD can establish the order of dependence between a dependent variable and independent variables, each of MRA and BAYD could serve as a pioneering dimension-reduction tool in data mining. In the case study, since the nonlinearity of the studied problem is strong, the nonlinear algorithm BAYSD can serve as a pioneering dimension-reduction tool, but the linear algorithm MRA cannot.

(d) Since the nonlinearity of the case study is strong, BPNN and R-SVM are not applicable though they are nonlinear algorithms, whereas other two nonlinear algorithms C-SVM and BAYSD are applicable, indicating the nonlinear ability of C-SVM and BAYSD is higher than that of BPNN and R-SVM.

(e) Comparing the two applicable algorithms C-SVM and BAYSD for this case study, it is seen that $\bar{R}^*$(%) of C-SVM is less than that of BAYSD; BAYSD can serve as a pioneering dimension-reduction tool, but C-SVM cannot; it is easy to code the BAYSD program whereas it is very complicated to code the C-SVM program, so BAYSD is a good software for this case study when C-SVM is not available.

REFERENCES

[1] Le Maitre, R. W. (1984). A proposal by the IUGS subcommission on the systematics of igneous rocks for a chemical classification of volcanic rocks based on total alkali silica (TAS) diagram. *Australian Journal of Earth Science*, *31*(2), 243-255.

[2] Qiu, J. X. (1991). Brief introduction of a classification of volcanic rocks recommendations of the IUGS subcommission on the systematics of igneous rocks. *Geoscience*, *5*(4), 457-468.

[3] Zhu, Y. X., & Shi, G. R. (2013). Identification of lithologic characteristics of volcanic rocks by support vector machine. *Acta Petrolei Sinica*, *34*(2), 312-322.

[4] Shi, G. R. (2013). *Data mining and knowledge discovery for geoscientists*. USA: Elsevier Inc.

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