The Influence of Vibratory Excitation on the Oil Slug Mobilization in a Capillary Model

L. Dai[1*], G. Cheng[1], M. Dong[2], Y. Zhang[1]

[1]Industrial Systems Engineering, University of Regina, Regina, Saskatchewan, S4S 0A2 Canada

[2]Chemical and Petroleum Engineering, University of Calgary, Calgary, Alberta, T2N 1N4 Canada

liming.dai@uregina.ca

Abstract

This research investigates both experimental study and numerical simulation for the mobilization of residual oil under vibratory conditions. A capillary model is established to present pore structure and various vibratory stimulations are applied to the model. Through experimental studies, a relationship is built between the driving pressure and the system parameters such as drainage time, oil slug length, vibration frequency and duration. The numerical study shows the existence and development of a water film surrounding the oil slug during the oil slug mobilization. Results indicate that optimal vibration frequency and duration time can greatly increase the oil slug mobilization efficiency.

Keywords

Liquid Mobilization, Porous Media, Enhanced Oil Recovery, Slug Flow, Vibratory Excitation, Capillary

Introduction

The petroleum industry has implemented enhanced oil recovery (EOR) techniques and recognizes that oil production is generally improved by applying vibro-seismic techniques. The exact mechanism behind it is far from understood. While it has been demonstrated by numerous experimental and field test results that vibro-seismic stimulation can improve waterflooding efficiency during the residual oil recovery process in depleted reservoir [1], practical implementations of such techniques are at best trial and error. Often, vibration frequency on the order of mega Hertz in dynamic range is swept to mobilize residual oil. Such a wide range of vibration frequency, to some extent, reflects the lack of efficient and effective guidance in utilizing such techniques. Thus, it is theoretically and practically significant to carry out systematic research on the flow behavior of residual oil during the mobilization process under vibratory stimulation.

Ashchepkov [2] conducted a series of experiments on sandstone cores with different permeability. Vibratory stimulations at frequencies of 30, 60, 100, 200, and 400 Hz and with amplitudes of 0.6 and 0.4 μm were applied to these cores. Significantly faster imbibition of water and oil drainage was observed after the vibratory treatment. Pogosyan [3] studied the influence of vibratory stimulation on single-phase permeability on a synthetic core made from quartz sand and marshallite. It was observed that the permeability and drainage of the core were increased by vibration; however, the nature of the vibration stimulation was not specified.

Simkin and Surguchev [4] conducted a spectrum of experiments including gravity segregation, oil-displacement, and imbibition on a sandpack. The gravity segregation experiments showed that when the sandpack was vertically inverted without vibration, it took approximately 12 days to achieve the new gravity equilibrium; however, the new gravity equilibrium was achieved in only about 2 hours under a vibratory stimulation of 120 Hz and amplitude of 0.01cm. It was observed in the oil-displacement tests that, with vibratory stimulation, an oil recovery of 75% was achieved while without vibration, the observed oil recovery was approximately 45%. In the imbibition tests conducted on a sandstone core after 352 hours of imbibition, oil recovery of 56% was achieved without vibration whereas under vibratory stimulation, the oil recovery reached 96%.

Beresnev's research group from Iowa States University has been working in this area for over 10 years and has made a great contribution in the effects of vibration on oil mobilization[1, 5-9]. Through theoretical research, they

2168-5517/16/01 001-13, © 2016 DEStech Publications, Inc.
doi: 10.12783/jpsr.2016.0501.01

found that vibratory stimulation could remarkably decrease the value of the pressure gradient required to mobilize the entrapped oil slug. They also considered a 2-D etched glass micromodel with a 50×50 square lattice of circular "pores" connected by straight channel "throats". The experimental results demonstrated that for fixed acceleration amplitude, TCE was displaced faster by water injection as the frequency decreased from 60 to 10Hz. For the fixed vibration frequency, the TCE mobilization was enhanced as the acceleration amplitude increased from 0.5 to 5.0m/s2. Their latest results reported in 2011 described a sinusoidally constricted capillary model to simulate pore-throat in porous media and studied the impact of vibration excitation on oil slug mobilization; their results shew that vibration had positive effects on the process of oil slug overcome pore-throat. Also, the lower frequency of vibration works better compared with higher frequency ones.

Roberts [10] conducted experiments by applying compression vibration to a sandstone core in the axial direction. It was observed that oil flow was enhanced at frequencies between 25 to 100 Hz, whereas no significant changes in permeability were observed. Experiments were conducted by Wang [11] and Spanos [12] on vertical columns of sandpack with oil flowing through it. Unlike other experiments, pressure pulses at a frequency of about 1 Hz were applied to the sandpack. The experimental results showed a significant increase in the flow rate. Ma [13] investigated the effect of vibration on core permeability (32 cores with diameter: 2.5 cm, length: 3 to 7.5 cm under translational vibration of 0.05 to 0.2 g force at frequencies of 0 - 30 Hz). It was observed that the permeability increased by 28% when the frequency was the eigen-frequency of the core, whereas permeability actually decreased at other frequencies. However, they failed to define the core's eigen-frequency in their paper.

The research group at the University of Regina has been working on the effect of vibration excitation on oil slug mobilization in capillary model. Dong [14] studied the pressure drop of oil slugs, mobilized by very low and constant injection rates in a water-filled capillary tube. As found in their research, to mobilize the oil slug in straight capillary tube, the driving pressure must be large enough to overcome the resistance of the capillary wall as well as the viscous pressure. Dai and Zhang [15] illustrated four different flow phenomena during the process of oil slug mobilization in a capillary; their results indicate that the development of water film is crucial in oil slug mobilization, and the flow phenomena are dominated by the oil slug length as well as the water injection rate. Later on, Dai and Wang [16] conducted a numerical research on the oil slug mobilization in an axisymmetric capillary tube. The results confirmed the existence of water film between the oil slug and the tube wall. Fan [17] studied the effect of vibratory stimulations on a flowing slug. For an oil slug flowing in a capillary tube under a constant driving pressure, the results showed the existence of an optimum vibration frequency, under which the oil slug achieved its maximum speed. Cheng [18, 19] also conducted experiments to investigate the impact of vibratory stimulation on the mobilization of oil slugs in a capillary tube. It was found that within a certain frequency range, the driving pressure that was required to mobilize the oil slug decreased with the increase of the vibration frequency.

In this research, a customized test bed is presented to study oil slug mobilization in porous media along with the experimental results that provide insights on the impact of water drainage time, vibration frequency, and vibration duration on a required oil slug mobilization pressure. The results obtained in the present research contribute to the efforts of revealing mechanisms of the vibratory stimulation technique in EOR.

Experimental Test Bed and Approach

Experiment Test Bed

The layout of the experimental test bed is depicted in Figure 1. The experimental setup consists of the pore structure model, the vibration generation system and the data acquisition system.

For the sake of clarity and simplicity, a capillary model is used to present a branch of pore structure underground. The core component of pore structure model is a uniform circular glass capillary tube, which is 1000 mm long with an inner diameter of 1.5 mm (PYREX Brand Glass Tubing). Two Plexiglas chambers (Chamber A and Chamber B) are attached to fix the tube. A pulse-free constant syringe pump (Model 341A, Sage Instrument) is connected to Chamber B to provide a constant flow rate in the capillary tube. On the other side, a water reservoir is connected to

Chamber A to collect the outflow liquid.

The data acquisition system is used to collect and record the pressure differences across the capillary tube throughout the oil slug mobilization process. The core device of the data acquisition sub-system is a highly sensitive pressure transducer (DP103, Validyne Engineering) with a resolution as low as 0.014 mmH2O. The signals from the transducer are collected and converted into digital signals by a data acquisition board (CIO-MINI37, National Instruments) connected directly to the carrier demodulator. The digital signals generated by the data acquisition board are then transformed into computer data and are processed by a data processing software (DASYLab V8.0, National Instruments) installed on the computer. The collected data are stored as raw experimental data which can be plotted to visualize the pressure profile.

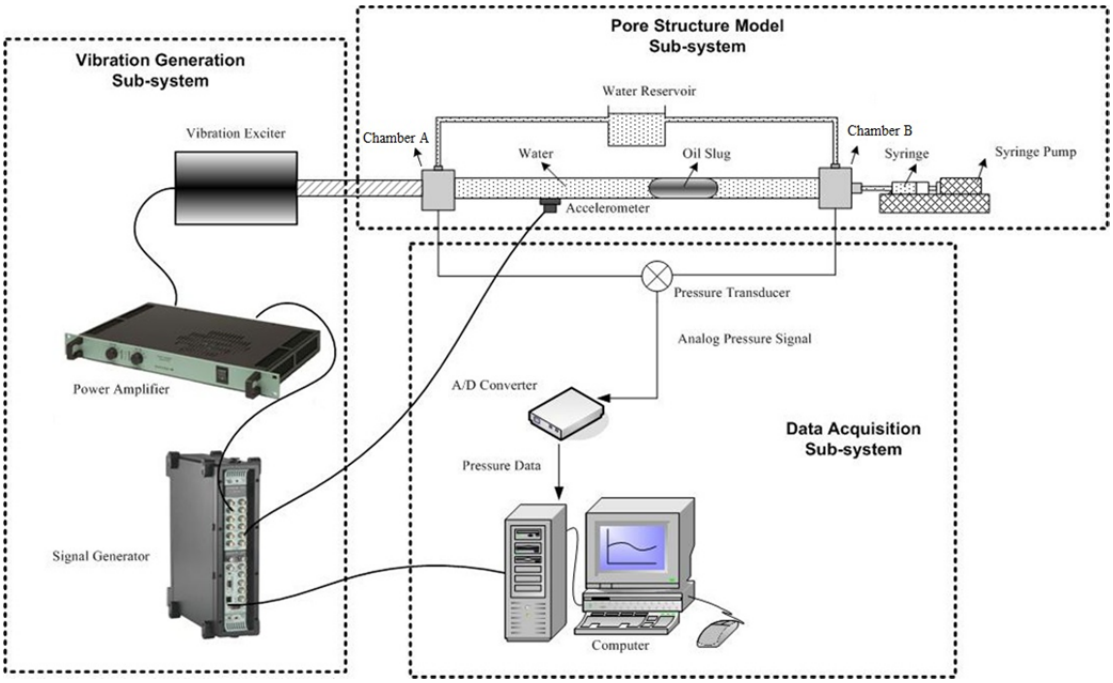

FIGURE 1: EXPERIMENTAL TEST BED.

The vibration generation system consists of a vibration exciter (Bruel & Kjaer, Type 4808), a power amplifier (Bruel & Kjaer, Type 2706) and a signal generator (Bruel & Kjaer, Type 3560C). An accelerometer (DeltaTron, Type 4398) is used to monitor the vibration amplitude of the capillary tube. The vibration generation system is controlled by a software (Bruel & Kjaer, Pulse 8.0), which is also used for data acquisition. The vibration system is capable of generating vibration signals with frequencies ranging from 5 Hz to 10 KHz. Waveforms of the vibration signals can be sinusoidal, impulsive and so forth. Sinusoidal signals are used in this research to study the effect of vibratory stimulation on oil slug mobilization.

Experimental Material Properties and Conditions

The oil sample used in this research is standard S60 in conformity with the ASTM Oil Standard. In order to minimize the adverse effect on the experiment results caused by impurities, DI water is employed in oil slug mobilization. Experiments are conducted at room temperature (22 ± 0.5˚C). The oil-water interfacial tension (IFT), γow, of the oil sample is measured to be 27 dyne/cm using the pendant drop method. Thus, the capillary pressure is calculated to be 2γow/Rt = 7.2 mmH2O with a contact angle of 0°. At room temperature, the viscosity of the S60 oil sample is 148.0 cp and the water viscosity is 1.0 cp. Densities of the oil and water are 0.87 and 1.00 g/cm3, respectively.

Experimental Procedures

The purpose of the current research is to study the effects of both the vibration frequency and duration on the mobilization of an oil slug in the capillary tube, respectively. The following experimental procedures are carefully considered to reach the above goal.

1) A capillary tube is firstly cleaned by Varsol, acetone, DI water orderly; and then blows dry with air for an hour to ensure the inner surface is completely clean and dry.
2) The clean tube is fixed by Chamber A and B, the whole system is then filled with DI water, it is necessary to make sure there is no air bubble into the flow system.
3) An oil slug with desired length is injected to the capillary tube through Chamber B. It should be noted that after the oil slug is injected into the capillary tube, the needle connected to the syringe must be pulled out of the tube very slowly and steadily so that the oil slug would not break due the motion of the needle. Such a slow motion of the needle also prevents the inner surface of the capillary tube from being scratched by the sharp point of the needle.
4) The flow system is then standing there without any external pressure and waiting for the drainage time. The drainage time is defined as the duration of time for which the oil slug stays still inside the capillary tube. During the drainage time, the water film exists between the oil slug and tube wall drain out due to the interfacial tension between oil and water, such that the oil slug contacts with the tube wall directly. Water film drainage is an important process and it will be discussed in section 4.1 in detail.
5) When the oil slug is set ready, after desired hours of drainage, the capillary tube is subjected to the stimulation of external vibration. A vibratory stimulation ranging from 5 to 350 Hz is applied to the capillary tube containing the oil slug, and the acceleration amplitude is 8 m/s2.
6) After a desired time of vibration, the exciter is shut down and water is injected immediately by the syringe pump. At the meantime, the pressure difference across the tube is measured and recorded during the water injection.
7) When the oil slug is observed to flow in a stable condition and the pressure drop trend becomes smooth, then the water injection is stopped. This is considered as an entire cycle of experiment.

Results and Discussion

Influence of Water Film Drainage on Oil Slug Mobilization

1) Experimental Research of Influence of Water Film Drainage on the Oil Slug Mobilization

Firstly, we investigate the impact of water film drainage on the pressure needed for oil slug mobilization. For a moving oil slug, research has shown that when the capillary tube radius is small enough such that gravity is negligible, there will be a film of liquid with uniform thickness between the oil slug and the capillary wall [20]. When an oil slug is initially injected into a capillary tube full of water, it is considered as a moving oil slug. At this stage, a water film will be generated once an oil slug is injected to the capillary tube. When the oil slug stops moving and remains still inside the capillary tube, the water film will be partially or fully drained out under surface tension of the oil slug eventually.

There are many factors that affect the water film drainage for oil slugs such as temperature, oil slug length and material properties of the oil. In this research, extensive work is focused on investigating the relationship between the driving pressures required for the mobilization of oil slugs of different lengths and with different durations of drainage time.

A set of mobilization experiments were conducted on an oil slug inside the capillary tube with a range of drainage times between 2 to 38 hours, depending on the oil slug length; generally speaking, the shorter oil slug is, the shorter time it needs to drain out the water film between the oil slug and the tube wall. In the previous research, Dai [16] shows that the value of the maximum driving pressure is mainly dependent on the oil slug length and the water injection rate does not have significant influence on the value of the maximum driving pressure. Thus, the water injection rate is fixed at 0.35 ml/hr in this research. For each drainage time, the maximum driving pressure (pressure needed for slug mobilization) was collected. This set of experiments was then performed for several slug lengths: 4 mm, 8 mm, 20 mm, 30mm and 65 mm. It should be mentioned that longer drainage time (up to 50 hours) has been applied to 65 mm oil slug in order to make sure the water film is maximally drained out.

Figure 2 shows the trend of a typical pressure profile summarized from the oil slug mobilization experiment. The driving pressure applied to the slug was proportionally increasing over experiment time until the slug began to mobilize. This pressure was recorded as the maximum driving pressure needed for slug mobilization. Subsequently, less pressure was needed since the slug was already mobile.

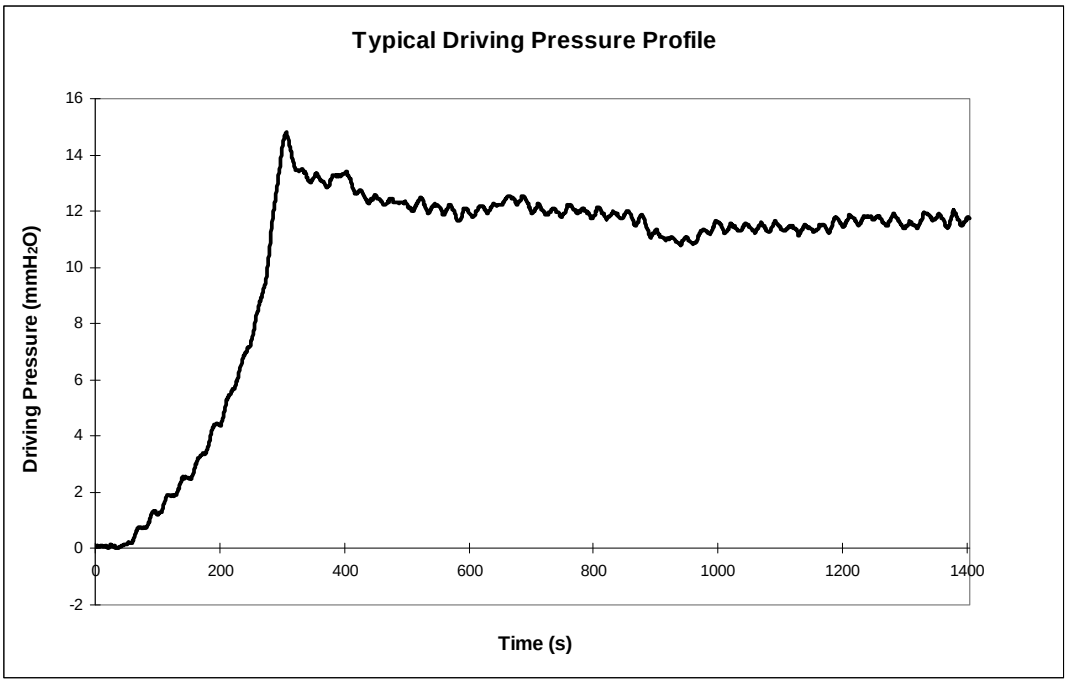

FIGURE 2: A TYPICAL DRIVING PRESSURE PROFILE

Similar pressure profiles were obtained for other drainage times for specific slug lengths. The relationship between the maximum driving pressure and drainage time is plotted in Figure 3. As expected, the shorter the drainage time is, the lower the maximum driving pressure is required for oil slug mobilization. This confirms our earlier results on the driving pressure required to mobilize the oil slug [19]. The decrease in maximum driving pressure with decreasing drainage time can be interpreted in the following way: when the drainage time is shorter, there is still a considerable amount of water film between the oil slug and capillary wall, which makes mobilization easier, therefore requiring less driving pressure. On the other hand, when the drainage time is longer, there is less amount of water film existing between the oil slug and capillary wall, which makes the mobilization of the oil slug more difficult, therefore requiring more driving pressure.

Although the maximum driving pressure increases with drainage time, the increasing rate decreases significantly beyond a certain point. Take the maximum pressure curve of 4 mm oil slug as example; the maximum driving pressure is almost constant between 13 to 38 hours of drainage time. Intuitively, when the drainage time is long enough depending on the oil slug length (for example, 13 hours for a 4 mm oil slug), the water film will be drained to the utmost extent such that there would be no significant difference from the maximum driving pressure needed for longer drainage times.

Meanwhile, Figure 3 compares the maximum driving pressure recorded for oil slugs with different lengths. As expected, the maximum driving pressure needed to mobilize the oil slug decreases as the oil slug length increases. While all curves exhibit very similar trends. All of the maximum driving pressures increase with increasing drainage time. However, the slope of the increasing driving pressure reduces notably as the oil slug length increases. Also, as one may notice from the figure, the rate of increase in maximum driving pressure for longer oil slugs is much less than that for shorter ones. This implies that the water film or the amount of residual water between the oil slug and the solid wall plays an important role in mobilizing the oil slug in capillary. One may also conclude that the maximum driving pressure required to mobilize an oil slug is less sensitive to drainage time for longer oil slugs. Our results on the relationship between slug length and maximum driving pressure provide further insights on the level of difficulty in slug mobilization for porous media with different amount of surface contact with slugs.

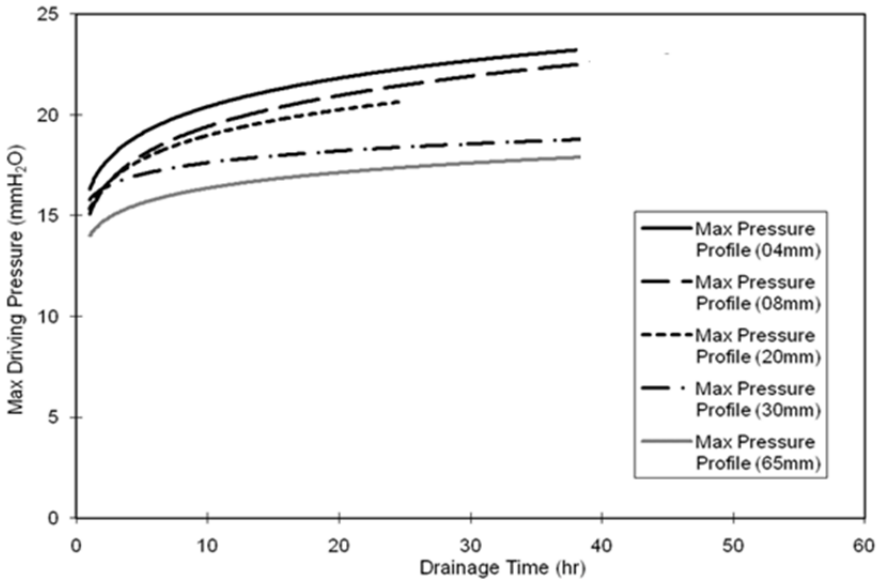

FIGURE 3: MAXIMUM DRIVING PRESSURE VS. DRAINAGE TIME FOR OIL SLUGS OF DIFFERENT LENGTHS

2) Numerical Research of Influence of Water Film Drainage on the Oil Slug Mobilization

Numerical simulation results have been provided for better understanding of the effect of drainage time on the oil slug mobilization. A powerful computational fluid dynamics software package ANSYS-FLUENT has been employed to simulate the experimental observation about the drainage of the water film. The most important task in a numerical simulation is the establishment of a proper model. A horizontal capillary tube 80 mm long with an inner diameter of 1.5 mm is established based on the geometric dimensions of the capillary tube used in actual experiments; inside the capillary tube, two immiscible fluids with different physical properties are considered: a 2 mm long oil slug and in a tube filled with water. The properties of the oil slug and water are the same as given in section 2.

An inner-surface tracking method namely volume of fluid (VOF) algorithm is applied here to simulate the oil-water flow in capillary tube [21]. In the VOF model, the surface position is determined by fraction function (Eq. 2, 3, 4). Then the Navier-Stokes equation (Eq. 1) is applied based on the results of this fraction function.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = 0 \tag{1}$$

$$\frac{\partial}{\partial t}(\rho u) + u \cdot \nabla u = \nabla p + \mu \nabla^2 \mathrm{u} + \mathrm{f} \tag{2}$$

$$\rho = \alpha \rho_1 + (1-\alpha)\rho_2 \tag{3}$$

$$\mu = \frac{\alpha \rho_1 \mu_1 + (1-\alpha)\rho_2 \mu_2}{\alpha \rho_1 + (1-\alpha)\rho_2} \tag{4}$$

whereϱ , ϱ_1 and ϱ_2 are the average density, density of phase 1 and density of phase 2, respectively.μ, $\mu_1$and μ_2 are the average viscosity, viscosity of phase 1 and viscosity of phase 2, respectively, p is the pressure, u is the velocity, f is the body force per unit volume and α is the volume fraction of phase.

Figure 4-(a) illustrates the oil slug in the initial state and Figure 4-(b) to (d) give the visualized contours of the diminishing water film. As shown in the figure, the thickness of the water film decreases, as the oil slug remains stationary inside the capillary tube. Given enough time, the water film will eventually be drained out from the space between the oil slug and the tube wall, then the oil slug will contact with the tube wall directly, which makes it difficult to be mobilized because, in addition to the capillary force, there is also resistance between the tube wall and the oil slug due to the viscous force.

The phenomenon appears in the numerical simulation can explain the experimental result, that the longer

drainage time is, the higher pressure it needs to mobilize the oil slug, because of the increasing of the resistance force.

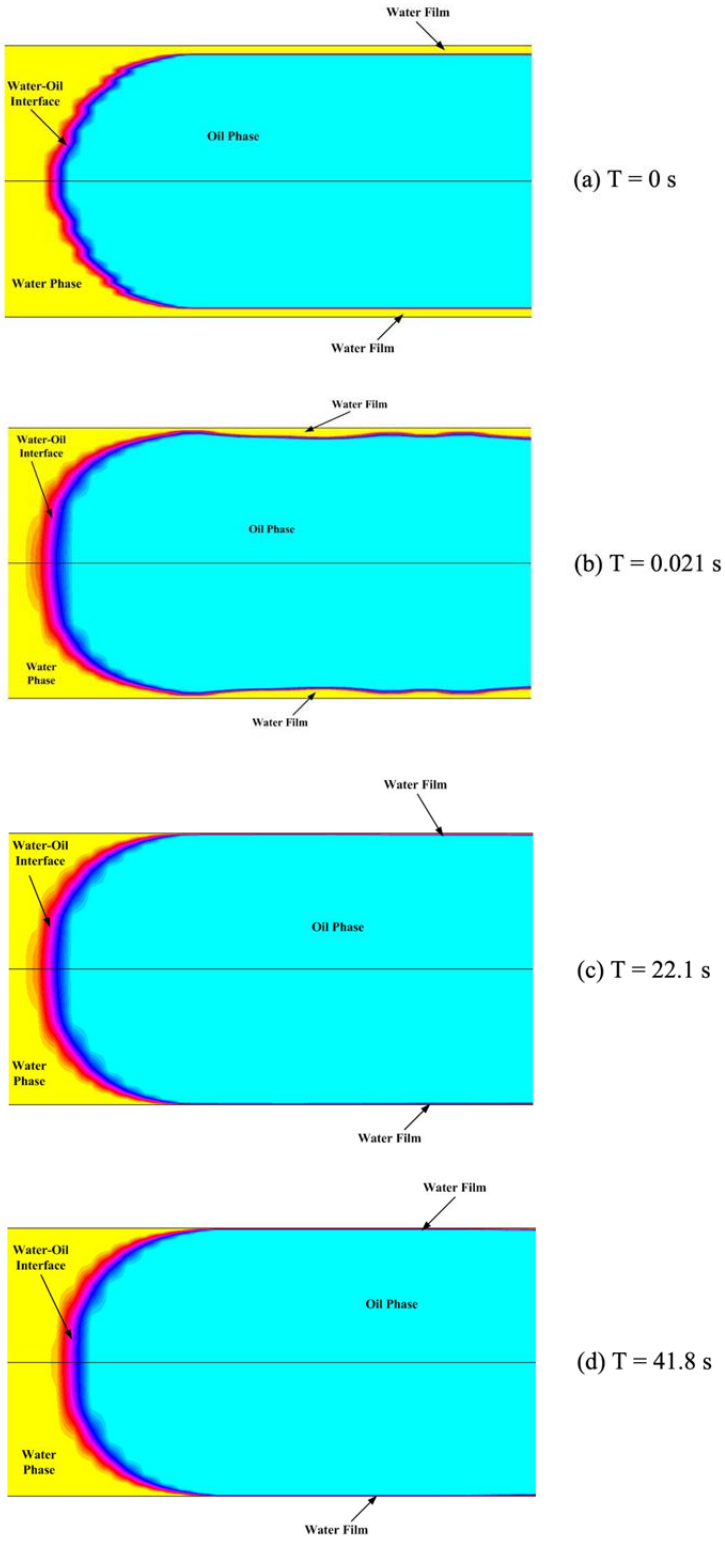

FIGURE 4: DRAINAGE OF WATER FILM

Influence of Vibration Excitation on Oil Slug Mobilization

The main focus of this research is to study the effects of vibration on oil slug mobilization in a capillary model. The influence of both vibration frequency and vibration duration time is analyzed in detail.

1) Influence of Vibration Frequency on Oil Slug Mobilization

Firstly we fix the oil slug length (25 mm) and drainage time (24 hours) in this experiment to explore the impact of vibration frequency on the maximum driving pressure.

After treating the pore structure model under vibration stimulations at different frequencies ranging from 5 to 350 Hz and for the same period of time (30 minutes), the mobilization of the oil slug in straight tube was initiated by injecting water into the capillary tube at injection rate equals to 0.34 ml/hr. The pressure difference between the two sides of the capillary tube was measured with respect to time. For an oil slug with a length of 25 mm and a drainage time of 24 hours, pressure profiles corresponding to different vibration frequencies are plotted in Figure 5.

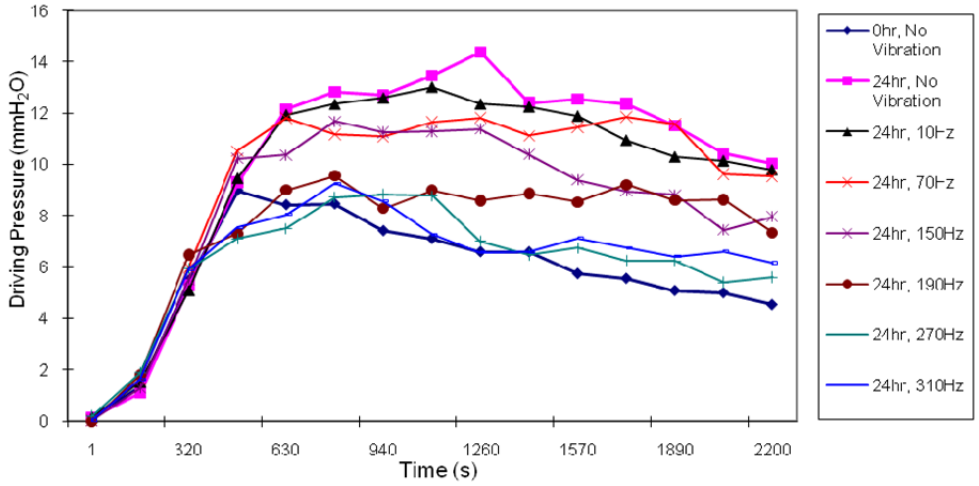

FIGURE 5: COMPARISON OF PRESSURE PROFILES FOR DIFFERENT VIBRATION FREQUENCIES ON A 25 MM OIL SLUG

As can be seen from Figure 5, the driving pressure required for mobilizing the oil slug decreases as the frequency of the vibratory stimulation increases. When the frequency reaches 270 Hz, the pressure profile obtained at this frequency almost overlaps with the lower bound pressure curve, i.e., the pressure profile obtained under the condition of zero drainage time and without vibration. For pressure profiles obtained from vibratory stimulation at frequencies higher than 270 Hz, they are so close that they almost overlap. The effect of mobilizing a 25 mm long oil slug in a capillary tube for 24 hours after vibratory stimulations for the duration of 30 minutes, when the frequency reaches 270 Hz or higher, can be deemed as equivalent to the effect of mobilization with zero drainage time. Figure 6 shows the maximum value of driving pressure at different vibration frequencies.

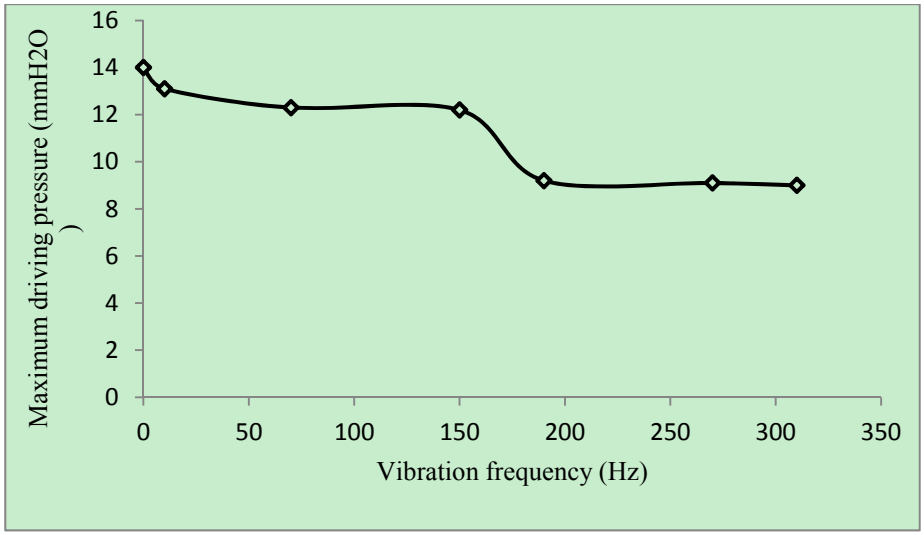

FIGURE 6: MAXIMUM DRIVING PRESSURE FOR DIFFERENT VIBRATION FREQUENCIES (25 MM OIL SLUG)3.2.3 INFLUENCE OF DURATION OF VIBRATION STIMULATION ON OIL SLUG MOBILIZATION

The experimental results presented in Figure 5 and Figure 6 can be summarized as follows:

1. The vibration excitation evidently shows positive influence on the mobilization of oil slugs in the capillary tube. The pressure shows two stages with increasing vibration frequency as shown in Figure 6, when all the other conditions are fixed. The effects of the vibration frequency on the maximum pressure can be seen as the following.
   - When the vibration frequency is equal or less than 150 Hz, the maximum driving pressure for mobilizing the oil slug is dropped by about 14% from that without vibration excitation;
   - When the vibration frequency is in the range of 190 Hz to 310 Hz, the maximum driving pressure is reduced 32% from the case without vibration excitation, or 18% of further reduction from the case above.
2. There exists an optimum vibratory frequency at which the best effect of the vibration applied on improving mobilization can be achieved. It means the effect of the optimum frequency can be made very close to the effect achieved when the drainage time is zero.

In addition to vibration frequency, the duration of vibration also shows impact over the driving pressure needed for slug mobilization. In this experiment, the vibration amplitude is 0.002 m and the frequency is fixed at 10 Hz. The vibration durations are ranging from 10 to 120 minutes. After treating the pore structure model under vibration stimulations with varying periods of time, the mobilization of the oil slug in straight capillary tube is initiated by injecting water into the capillary tube at injection rate equals to 0.34 ml/hr. Figure 7 shows a comparison of the pressure profiles obtained on a 43 mm long oil slug with drainage time of 24 hours under vibration stimulations of different durations.

As shown in Figure 7 the driving pressure required to mobilize the oil slug decreases as the duration of the vibratory stimulation increases. The pressure profile almost overlaps with the lower bound (the pressure profile obtained under the condition of zero drainage time and without vibration) for the 43 mm long oil slug when vibration lasted 120 minutes. As per the experimental results obtained, the vibration of sufficiently long duration is preferred for mobilizing the oil slug with minimum driving pressure. Figure 7 shows the maximum value of driving pressure at different vibration duration time.

The results illustrated in Figures 7 and 8, demonstrate that vibration duration plays an important role in the mobilization of oil slugs in a capillary tube. It can be concluded that the best effect of vibration stimulation on mobilization can be achieved if the vibration duration used is sufficiently long. It should be noticed, however, the longer is the vibration duration, the more efforts are needed for generating and maintaining the vibration. For practically and economically sound results, the sufficiency of the vibration duration time needs to be determined.

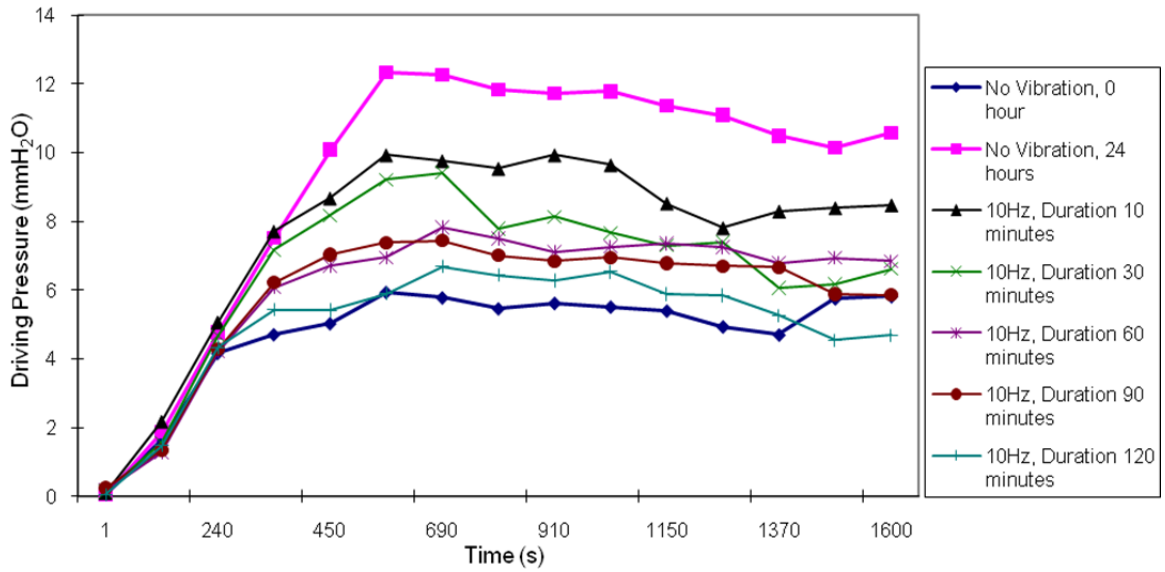

FIGURE 7: COMPARISON OF PRESSURE PROFILES FOR DIFFERENT VIBRATION DURATIONS ON A 43 MM OIL SLUG

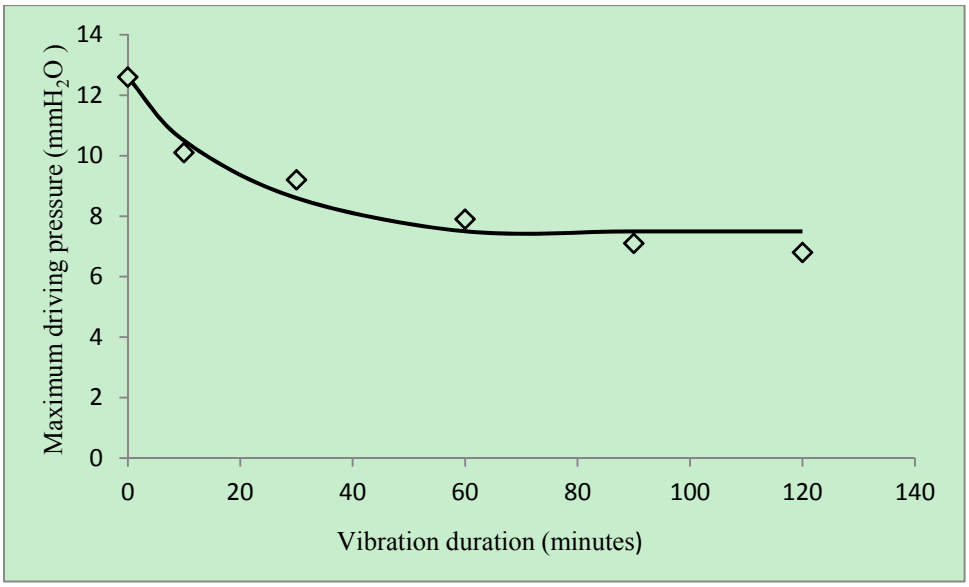

FIGURE 8: MAXIMUM DRIVING PRESSURE AT DIFFERENT VIBRATION DURATION TIME (43 MM OIL SLUG)

Since the driving pressure needed for oil slug mobilization decreases under the influence of vibratory stimulation, and, considering the experimental results, it is logical to hypothesize that the mechanism of how vibratory stimulation improves the mobility of an oil slug lies in the fact that vibration facilitates the development of water film between the oil slug and capillary wall. The higher the frequency is, the more developed the water film becomes, when the duration of the vibration is fixed. Likewise, the longer the duration of vibration stimulation is, the more developed the water film becomes, when frequency of the vibration is fixed.

2) Numerical Study on the Influences of Vibration Excitation on the Water Film Development

In the previous section, the positive effect of vibration excitation on oil slug mobilization is demonstrated experimentally. An optimal mobilization can be achieved if an external vibratory excitation of proper frequency is applied, under the conditions of the experiments. Also, the optimal oil slug mobilization is a case in which a water film is completely surrounding the oil slug without any drainage time. Therefore, it is logical to reach the hypothesis that external vibratory simulation positively affects the mobilization of the oil slug in straight capillary tube by facilitating the development of a water film between the oil slug and the capillary tube wall. However, it is difficult to verify such a hypothesis experimentally due to the limitation of experimental equipment. Numerical simulation therefore becomes an effective and efficient alternative for the verification of this hypothesis.

Same as described in section 3.1.2, ANSYS-FLUENT is applied to simulate the experimental procedures without water injection. In the numerical simulation, an oil slug of 20 mm in length is initially placed in the capillary tube without water film surrounding it. A user defined function (UDF) is then employed to provide a longitudinal tube movement, which follows sinusoidal change in the boundary condition. This generates a sinusoidal excitation similar to the flow system used in the experiments.

Figure 9 shows the numerical results. As shown in the figure, the changes in the shape of the oil slug before and after applying vibration excitation into the system. The oil slug is originally placed in the tube with two semi-circle shapes at the end and the main part of the oil slug is contacted with the tube wall. When subjected to external vibratory stimulation, the bulk of the oil slug remains in its initial position. However, as vibratory excitation continues, the two ends of the oil slug begin to change their shapes and become slender. Meanwhile the water begins to seep into the area between the oil slug and the capillary tube wall and a thin water film is formed. Also, although the water film is formed, the thickness of the water film is not uniformly distributed.

It is clearly demonstrated in the numerical simulation results that under the influence of external vibratory stimulation, a layer of water film is developed between the oil slug and the capillary wall. It is this water film that facilitates subsequent mobilization of the oil slug using water injection. Therefore, the numerical result

supports the hypothesis that vibratory stimulation can facilitate mobilization of the oil slug by developing a water film between the oil slug and the capillary wall. It is this water film that makes subsequent mobilization of the oil slug by water injection easier with a lower driving pressure.

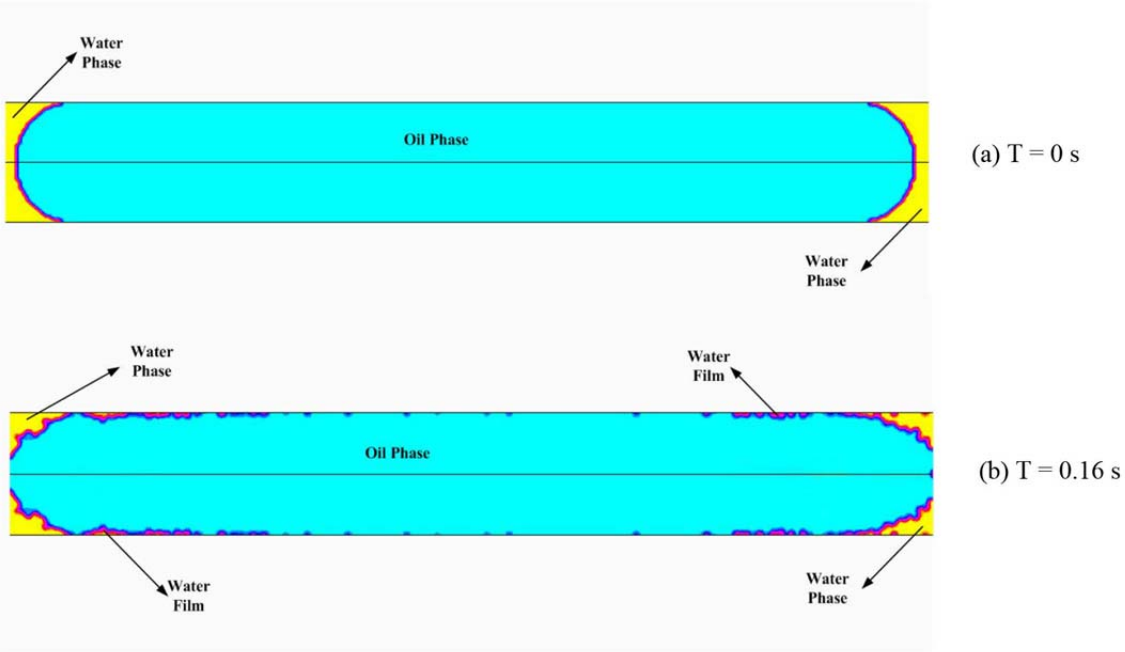

FIGURE 9: DEVELOPMENT OF WATER FILM WITH VIBRATION EXCITATION

Conclusions

The impact of vibration on the amount of water pressure needed to displace residual oil in porous media is not well understood. To study the impact of vibration frequency and duration on oil slug mobilization, a series of experiments on a custom designed experimental test bed are carried out. This is a systematic laboratory exploration of the effect of vibratory stimulation on the driving pressure needed for oil slug mobilization. While the experimental model is idealized, our results provide useful insights on the mechanism of mobilizing oil in porous media under vibratory stimulation.

The following conclusions can be drawn from the results of the research:

1. For an oil slug with certain length, the maximum driving pressure increases with drainage time, the increasing rate decreases significantly beyond a certain point; for oil slugs with different lengths, the maximum driving pressure required to mobilize an oil slug is less sensitive to drainage time for longer oil slugs.
2. The frequency of the vibratory stimulation plays an important role in oil slug mobilization. Given an oil slug initially in a stationary state, the maximum driving pressure needed to mobilize the oil slug decreases as the frequency increases when the other factors are held constant. When the frequency becomes sufficiently high, the vibratory stimulation can achieve the mobilization effect very close to that achievable under the ideal condition of zero drainage time.
3. The duration of the vibratory stimulation also plays a vital role in the effect of oil slug mobilization, as revealed in the experimental analysis. Given the oil slug is initially in a stationary state, the longer the duration of the vibratory stimulation is, the smaller the maximum driving pressure is needed to mobilize the oil slug when vibration amplitude and frequency are fixed. When vibratory stimulation lasts for a sufficiently long time, it can achieve the same mobilization effect as it would be under a condition of zero drainage time.

4. Numerical research is employed to create simulations of the processes of both the water film drainage and the effect of vibration on the water film development. The numerical results conclude the following:
   (1) During the drainage process, the water film between the oil slug and the tube wall drains out due to the surface tension of oil slug.
   (2) For a stationary oil slug, the vibration excitation has positive effect on the development of water film between the oil slug and the tube wall.

By using driving pressure as an indicator of oil slug mobility, detailed studies of the impact of drainage time, slug length, vibration frequency, and vibration duration on the maximum required driving pressure are presented. This research makes contribution to revealing the mechanism of the vibratory mobilization of oil in porous media.

ACKNOWLEDGEMENTS

The authors acknowledge with thanks the support of research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Petroleum Technology Research Centre (PTRC).

REFERENCES

[1] Bresenev, I.A., and Johnson, P.A., Elastic-wave Stimulation of Oil Production: A Review of Methods and Results. Geophysics, 59(6), 1994, pp. 1000-1017.

[2] Ashchepkov, Yu. S., Infiltration Characteristics of Inhomogeneous Porous Media in a Seismic Field, Journal of Mining Science, 25(5), 1989, pp. 492-496.

[3] Pogosyan, A. B., Simkin, E. M., Stremovskiy, E. V., Surguchev, M. L., and Shnirel'man, E. M., "Separation of hydrocarbon fluid and water in an elastic wavefield acting on a porous reservoir medium", Dokl. USSR Acad. Sci., Earth Sci. Sect., 307, 1989, pp. 575-577.

[4] Simkin, E. M., and Surguchev, M. L., "Advanced Vibroseismic Technique for Water Flooded Reservoir Stimulation, Mechanism and Field Test Results", 6th European IOR Symp., Stavanger, Norway, 1991.

[5] Iassonov, P. and Beresnev, I., A Model for Enhanced Fluid Percolation in Porous Media by Application of Low-frequency Elastic Waves, Journal of Geophysical Research, 108(B3), 2003, doi: 10.1029/2001JB000683.

[6] Bresenev, I.A., Vigil, R.D., Li, W., Pennington, W.D., Turpening, R.M., Iassonov, P., and Ewing R.P., Elastic waves push organic fluids from reservoir rock. Geophysical Research Letter, 32, 2005, L13303, doi: 10.1029/2005GL023123.

[7] Li, W., Vigil, R.D., Beresnev, I., Issonov, P., and Ewing, R., Vibration-induced Mobilization of Trapped Oil Ganglia in Porous Media: Experimental Validation of a Capillary-physics Mechanism, Journal of Colloid and Interface Science, 289, 2005, pp. 193-199.

[8] Bresenev, I.A., and Deng, W., Viscosity effects in vibratory mobilization of residual oil. Geophysics, 75(4), 2010, pp. 79-85.

[9] Bresenev, I.A., Gaul, W., and Vigil, R. D., Direct pore-level observation of permeability increase in two-phase flow by shaking. Geophysical Research Letter, 38(20), 2011, L20302, doi: 10.1029/GL048840.

[10] Roberts, P.M., Igor, B.E., and Ernest L.M., 2003, Elastic Waves Stimulations of Oil Reservoirs: Promising EOR Technology? The Leading Edge, 22(5), 2005, pp. 448-453.

[11] Wang, J., Dusseault, M.B., Spanos, T., and Davidson. B., Fluid Enhancement under Liquid Pressure Pulsing at Low Frequency, Paper 1998.016, 7th UNITAR, pp. 151-159.

[12] Spanos, T., Davidson, B., Dusseault, M.B., Shand, D., and Samaroo, M., Pressure Pulsing at the Reservoir Scale: A New IOR Approach, Paper 99-11, CSPG and Canadian Petroleum Society Joint Convention, Calgary, Alberta, 1999.

[13] Ma, J., Liu, W., Huang, H., Liu, S., Zhang, H., and Zhang, Y., Increasing the Waterflood Recovery Efficiency of Core by Mechanical Vibration, Journal of Xi'an Petroleum Institute, 1997, 12(4), pp. 19-21.

[14] Dong, M., Fan, G., and Dai, L., An Experimental Study of Mobilization and Creeping Flow of Oil Slugs, Transport in Porous Media, 80 (3), 2009, 455-467.

[15] Dai, L. and Zhang, Y., Experimental Study of Oil-Water Two-Phase Flow in a Capillary Model, Journal of Petroleum Science and Engineering, 108, 2013, 96-106.

[16] Dai, L. and X. Wang, A Numerical Study on Mobilization of Oil Slugs in a CapillaryModel with Level Set Approach. Engineering Applications of Computational Fluid Mechanics, 8 (2), 2014, pp. 422-434.

[17] Fan, G., 2006, A Pore Scale Experimental Study of Residual Oil Mobilization and Seismic Stimulation, M.A.Sc. Thesis, University of Regina.

[18] Cheng, G. and Dai, L., Influence of Thin Film Drainage between Oil Slugs and Capillary Wall on Initial Differential Pressure, University of Regina Graduate Student Research Conference, University of Regina, Regina, 2008.

[19] Cheng, G. and Dai, L., Vibratory Mobilization of Two-Phase Liquid Trapped in Capillary, Proceedings of ASME 2009 International Design Engineering Technical Conferences (IDETC), San Diego. 2009, DETC2009-87021.

[20] Bretherton, F.P., The Motion of Long Bubbles in Tubes. Journal of Fluid Mechanics, 14, 1961, 81-98

[21] ANSYS FLUENT User's Guide. Chapter 26: Modeling Multiphase Flows. 2011, ANSYS, Inc.

The Use of Artificial Neural Network and Support Vector Classification for Recovery Factor Prediction

Dawei Li*[1], Guangren Shi[2]

Research Institute of Petroleum Exploration and Development, PetroChina, P. O. Box 910, Beijing 100083, China

*[1]leedw@petrochina.com.cn; [2]grs@petrochina.com.cn

Abstract

Oil and gas recovery factor is a key parameter for oil industry, and it can be effectively predicted by appropriate data mining algorithms. In this paper, three regression algorithms and three classification algorithms have been applied to forecast recovery factor of oil and gas. The three regression algorithms are the support vector regression (SVR), the artificial neural network (ANN), and the multiple regression analysis (MRA), while the three classification algorithms are the support vector classification (SVC), the naïve Bayesian (NBAY), and the Bayesian successive discrimination (BAYSD). The purpose of this paper is to demonstrate how to select proper algorithms in three algorithms (SVR, ANN, MRA) for recovery factor regression and/or three algorithms (SVC, NBAY, BAYSD) for recovery factor classification. In general, when all these six algorithms are used to solve a real-world problem, they often produce different solution accuracies. Toward this issue, it has been proposed that a) when an algorithm is applied to a real-world problem, its solution accuracy is expressed with the total mean absolute relative residual for all samples, R(%), and b) result availability of a given algorithm application is applicable if R(%)<10, and inapplicable if R(%)≥10. A case study of recovery factor in 39 global oilfields has been used to validate the proposed approaches. This case study consists of two problems: regression and classification. For the regression problem, only ANN is applicable since its R(%) value is 5.89, whereas SVR and MRA are inapplicable because their R(%) values are 68.9 and 38.4, respectively. For the classification problem, only SVC is applicable since its R(%) value is 0, whereas NBAY and BAYSD are inapplicable because their R(%) values are 24.7 and 34.5, respectively. From this case study, it is concluded that the preferable algorithm is ANN for recovery factor regression, while the preferable algorithm is SVC for recovery factor classification.

Keywords

Data Mining; Support Vector Regression; Artificial Neural Network; Multiple Regression Analysis; Support Vector Classification; Naïve Bayesian; Bayesian Successive Discrimination; Result Availability; Recovery Factor

Introduction

More and more data are being generated in subsurface geosciences (SUBG), and SUBG has already entered "big data" epoch for many years. For instance, there are about 50 large information systems in PetroChina, and there are about 590TB SUBG data in only one of them. At present, the major applications of SUBG data in these information systems are only storage and inquiry, which is far away from fully utilizing the values of these data assets (Li and Shi, 2015). Moreover, SUBG is facing more and more challenges. How to effectively utilize these "big data" to resolve some SUBG problems has become one of the study highlights for geoscientists. This has led to the generation of a promising frontier called *data mining* (DM) (Han et al., 2012), which is a preferable solution for this problem.

Since DM emerged from the late 1980's, many algorithms have been developed or introduced from other disciplines (such as mathematics, statistics, etc.) to DM (Shi, 2013). The major statistics algorithms in current DM can be classified into two groups: regression and classification.

Generally, there are three regression algorithms and three classification algorithms commonly used. The three regression algorithms are the support vector regression (SVR), the artificial neural network (ANN), and the multiple regression analysis (MRA), while the three classification algorithms are the support vector classification (SVC), the naïve Bayesian (NBAY), and the Bayesian successive discrimination (BAYSD).

In the recent years, regression and classification algorithms have seen enormous success in some fields of business and sciences, whereas the application of these algorithms to SUBG is still in fighting stage. This is because the

2168-5517/16/01 014-13, © 2016 DEStech Publications, Inc.
doi: 10.12783/jpsr.2016.0501.02

SUBG is very different from the other fields, with miscellaneous data types, huge quantity, different measuring precision, and many uncertainties to results. Moreover, most of the studied problems in SUBG are nonlinear.

In oil industry (a very important sector of SUBG), oil and gas recovery factor (i.e., the recovered ratio of the total oil and gas reserves) is a key parameter, as it directedly decides the total oil and gas production (Salehi et al., 2014). How to determine or predict oil and gas recovery factor has great significance for oil industry. However, there is almost no successful case study on this problem up to now, which was mainly because of lacking of related data and suitable techniques.

Based on the principles, methods and programs of DM introduced by Shi (2013), we have done many related researches of DM for oil industry (Shi et al., 2014; Shi, 2015a and 2015b; Zhang and Shi, 2015; Shi et al., 2015; Mi and Shi, 2015). For instance, Li and Shi (2015) took 13 samples as a case study of data mining in petroleum production, and adopted the six algorithms (SVR, ANN, MRA, SVC, NBAY, BAYSD) for forecasting formation flow capacity.

The purpose of this paper is to demonstrate how to collect enough data and select proper algorithms in three algorithms (SVR, ANN, MRA) for recovery factor regression and/or three algorithms (SVC, NBAY, BAYSD) for recovery factor classification. Using the data of 39 global oilfields, we applied the above six algorithms for forecasting both recovery factor and recovery factor classification, which differs much from other related articles to date, also being the major contribution of this paper.

In general, when all these six algorithms are used to solve a real-world problem, they often produce different solution accuracies. Toward this issue, it has been proposed that a) when an algorithm is applied to a real-world problem, its solution accuracy is expressed with the total mean absolute relative residual for all samples, R(%), and b) result availability of a given algorithm application is applicable if R(%)<10, and inapplicable if R(%)≥10. Here this threshold value (10) is taken due to the particularity of SUBG, while the threshold value (5) may be taken in other fields. Even in SUBG the threshold value (10) may be suitably adjusted based on a specific application.

The case study of recovery factor in 39 global oilfields from a commercial database has been used to validate the proposed approach. This case study consists of two problems: regression and classification.

Based on the above results availability of an algorithm for an application determined by its R(%), a flow chart of the six algorithms (SVR, ANN, MRA, SVC, NBAY, BAYSD) has been presented (Figure 1). Figure 1 illustrates the running path of each algorithm in the case study.

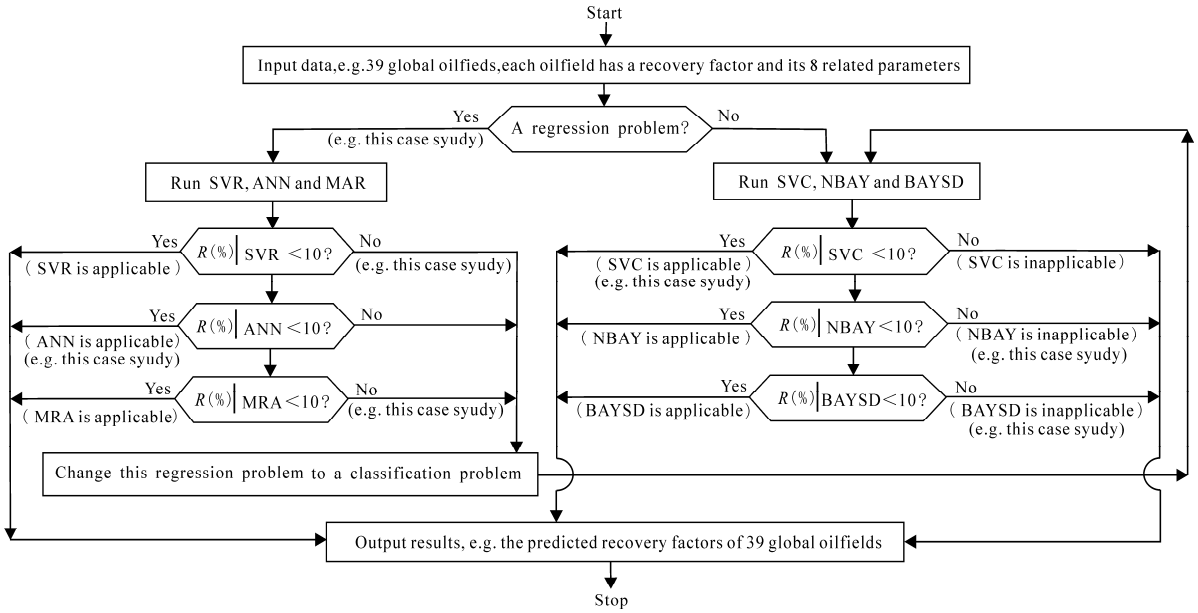

FIG. 1 A FLOW CHART OF THE SIX ALGORITHMS (SVR, ANN, MRA, SVC, NBAY, BAYSD)

Methodology

The methodology consists of the following two major parts: definitions commonly used by regression and classification algorithms; six algorithms (Shi, 2013).

Definitions Commonly Used by Regression and Classification Algorithms

The aforementioned regression and classification algorithms share the same sample data. The essential difference between the two types of algorithms is that the output of regression algorithms is real-type value and in general differs from the real number given in the corresponding learning sample, whereas the output of classification algorithms 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 algorithms (SVR, ANN, MRA, SVC, 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 parameters (x_1, x_2, …, x_m, y^*) and a set of observed values (x_{i1}, x_{i2}, …, x_{im}, y_i^*), with i=1, 2, …, n for these parameters. 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, …, m; and y_i^* is the observed value of the i^{th} learning sample. Equation 1 is the expression of learning samples.

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

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

where y=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 Eq. (1).

However, the principles of SVR and SVC 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 (x_{i1}, x_{i2}, …, x_{im}) into the fitting formula to get prediction values (y_1, y_2, …, y_n), respectively, so as to verify the fitness of an algorithm; 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}, …, 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 Eq. (1). Equation 3 is the expression of prediction samples.

In the six algorithms, only MRA is a linear algorithm whereas the other five are nonlinear algorithms, 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 algorithms are used, the following four types of residuals are defined.

The absolute relative residual for each sample, $R(\%)_i$ (i=1, 2, …, n, n+1, n+2, …, n+k), is defined as

$$R(\%)_i=\left|(y_i-y_i^*)/y_i^*\right|\times 100 \tag{4}$$

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). $R(\%)_i$ 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 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, $R_1(\%)$, is defined as

$$R_1(\%) = \sum_{i=1}^{n} R(\%)_i / n \tag{5}$$

where all symbols have been defined in Eqs. (1) and (4). $R_1(\%)$ is the fitting residual to express the fitness of learning process.

The mean absolute relative residual for all prediction samples, $R_2(\%)$, is defined as

$$R_2(\%) = \sum_{i=n+1}^{n+k} R(\%)_i / k \tag{6}$$

where all symbols have been defined in Eqs. (3) and (4). $R_2(\%)$ is the fitting residual to express the fitness of prediction process.

The total mean absolute relative residual for all samples, $R(\%)$, is defined as

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

where all symbols have been defined in Eqs. (1), (3) and (4). If there are no prediction samples, k=0, then $R(\%)=R_1(\%)$.

$R(\%)$ is the fitting residual to express the fitness of learning and prediction processes.

Six Algorithms

Each of the six algorithms (SVR, ANN, MRA, SVC, NBAY, BAYSD) is performed through two processes: learning process at first and then prediction process.

1) Learning Process

In the learning process, using learning samples expressed by Eq. (1), each algorithm constructs its own function $y=y(\boldsymbol{x})$. The methods of the six algorithms (SVR, ANN, MRA, SVC, NBAY, BAYSD) are described in Appendix A. It is noted that $y=y(\boldsymbol{x})$ created by ANN 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.

Substituting the values of m independent variables given by the n learning samples expressed by Eq. (1) into the constructed function $y=y(x)$, respectively, the result (y) of each learning sample for each algorithm is obtained. The mean absolute relative residual for all learning samples, $R_1(\%)$ defined by Eq. (5), for each algorithm is also obtained.

2) Prediction Process

Substituting the values of m independent variables given by the k prediction samples expressed by Eq. (3) into the constructed function $y=y(x)$, respectively, the result (y) of each prediction sample for each algorithm is obtained. The mean absolute relative residual for all prediction samples, $R_2(\%)$ defined by Eq. (6), for each algorithm is also obtained.

From learning process and prediction process, the total mean absolute relative residual for all ($n+k$) samples, $R(\%)$ defined by Eq. (7), for each algorithm is finally obtained.

Case Study: Recovery Factor in 39 Global Oilfields

This case study consists of two problems: regression and classification. The objective of this case study is to

calculate the recovery factor (RF, %) of oil and gas, and to determine the recovery factor classification (RFC) for oilfields that have few production data, which has practical value for predicting oil and gas production and evaluating their values in early E&P stage.

There are 39 global oilfield samples from a commercial database, and each sample contains 8 independent variables [x_1 = total production years, x_2 = current production stage, x_3 = current producing well count, x_4 = total well count, x_5 = original in-place oil equivalent (MMBOE), x_6 = EUR (Estimated Ultimate Recovery) oil equivalent (MMBOE), x_7 = production of cumulative oil equivalent (MMBOE), x_8 = production rate of current oil equivalent (BOEPD)] and one variable (y^* = RF or y^* = RFC). In the case study, among these 39 samples, 35 are taken as learning samples and 4 as prediction samples (Table 1) for the prediction of both RF and RFC, in which for RF using SVR, ANN and MRA, and for RFC using SVC, NBAY and BAYSD. It is noted that this RFC is figured out from RF by using the conversion rules given in Table 2.

TABLE 1 INPUT DATA FOR RECOVERY FACTOR IN 39 GLOBAL OILFIELDS

Sample type	Sample No.	Field No.	8 parameters related to y								y^*	
			x_1	x_2[a]	x_3	x_4	x_5	x_6	x_7	x_8	RF[b]	RFC[c]
Learning samples	1	F001	63	1	465	792	3200	385	311.7	16555	12.03	5
	2	F002	21	3	82	148	146	57	24	3884	10	5
	3	F003	46	5	193	308	1850	185	109	7646	32.59	3
	4	F004	58	3	63	288	405	132	126.8	2535	6.7	5
	5	F005	63	2	570	800	36840	2470	1452.4	103000	58.97	1
	6	F006	37	3	47	96	3900	2300	1830	63500	46	2
	7	F007	34	2	201	730	25000	11500	4394	447052	32.8	3
	8	F008	53	4	7	186	25	8.2	8.1	19	15.08	5
	9	F009	21	2	122	178	325	49	25	6353	6.19	5
	10	F010	59	1	2002	3632	13436	982	267.3	71470	37.18	3
	11	F011	44	3	153	430	347	129	122	4196	42.74	2
	12	F012	61	3	185	632	15667	7355	3650	335000	33.51	3
	13	F013	41	3	93	326	6900	3580	3697	211868	30.62	3
	14	F014	37	1	3200	4600	967	324	270	20580	36.26	3
	15	F015	54	1	145	350	60520	22530	7222	500000	24.39	4
	16	F016	37	4	179	437	353	128	116.6	1910	18.13	5
	17	F017	50	2	782	1218	41000	10000	7083	350000	20	5
	18	F018	45	1	100	155	5677	1410	744	105000	34.89	3
	19	F019	16	1	5	10	500	100	8.45	10000	65.85	1
	20	F020	97	5	838	3708	665	232	223	9078	25.64	4
	21	F021	112	3	9689	30000	4000	2634	2191	71550	15.6	5
	22	F022	88	1	630	1800	3900	1000	750	22000	14.9	5
	23	F023	9	1	24	42	1045	163	91	19514	32.68	3
	24	F024	48	1	88	112	314	46.8	37.9	4500	79.05	1
	25	F025	39	3	384	990	563	184	126	3746	35.36	3
	26	F026	118	3	10631	47000	4400	3478	3065	80019	34.37	3
	27	F027	55	4	60	173	110	38.9	34.8	551	63.5	1
	28	F028	74	4	51	331	483	166	165.4	390.03	11.12	5
	29	F029	42	6	60	86	1600	1182.7	789	80306	13.91	5
	30	F030	65	1	600	2500	1960	218	120	6900	33.33	3
	31	F031	52	6	95	226	611	85	68.6	5087	37.81	3
	32	F032	48	3	160	624	48000	16000	12782	522641	32.09	3
	33	F033	41	4	1710	3699	3420	1293	1164	54158	52.89	1
	34	F034	33	4	6	47	1153	370	360	2674	10.02	5
	35	F035	74	3	2221	6000	2800	1481	1340	26676	32.14	3
Prediction samples	36	F036	20	1	23	31	549	55	31	5900	(39.04)	(3)
	37	F037	74	1	2613	4121	28437	7565	5055	246441	(51.88)	(1)
	38	F038	44	4	33	73	140	45	41.8	198	(26.6)	(4)
	39	F039	25	1	327	643	637.3	94.1	59.8	11308	(14.77)	(5)

[a] In x_2, 1--primary rejuvenating, 2--secondary peak or plateau, 3--secondary decline, 4-- secondary mature, 5-- secondary rejuvenating, 6--tertiary peak or plateau.

[b] RF = the recovery factor (%) determined by the production data, number in parenthesis is not input data, but is used for calculating $R(\%)_i$.

[c] RFC = the recovery factor classification determined by Table 2, number in parenthesis is not input data, but is used for calculating $R(\%)_i$.

TABLE 2 RECOVERY FACTOR CLASSIFICATION

Recovery Factor	**RF (recovery factor) (%)**	**RFC (recovery factor classification)**
Very high recovery factor	>50	1
High recovery factor	40< RF≤50	2
Intermediate recovery factor	30< RF≤40	3
Low recovery factor	20< RF≤30	4
Very low recovery factor	≤20	5

TABLE 3 PREDICTION RESULTS OF RECOVERY FACTOR (RF) IN 39 GLOBAL OILFIELDS

Sample type	**Sample No.**	**RF**							
		y^*	**Regression algorithm**						
			SVR		**ANN**		**MRA**		
			y	$R(\%)_i$	y	$R(\%)_i$	y	$R(\%)_i$	
Learning samples	1	12.03	30.7558	155.659	11.6	3.88	18.7	55.1	
	2	10	33.2312	232.312	9.35	6.49	35.9	259	
	3	32.59	32.5433	0.143	32.1	1.67	28.9	11.4	
	4	6.7	31.0992	364.167	6.19	7.61	-.3.6	154	
	5	58.97	32.4701	44.938	58.1	1.56	33.2	43.7	
	6	46	31.862	30.735	47.2	2.52	42.5	7.62	
	7	32.8	32.3528	1.363	32.9	0.169	30.1	8.21	
	8	15.08	30.8125	104.327	15.2	0.629	22.6	49.7	
	9	6.19	30.8843	398.939	6.19	0.0000308	16.6	168	
	10	37.18	32.4168	12.811	36.4	2.05	29.3	21.3	
	11	42.74	33.1068	22.539	44.4	3.87	44.2	3.34	
	12	33.51	30.9235	7.719	35.0	4.53	29.7	11.4	
	13	30.62	32.1243	4.913	29.5	3.60	31.9	4.23	
	14	36.26	32.2356	11.099	35.6	1.88	30.7	15.4	
	15	24.39	31.7768	30.286	22.4	8.10	21.8	10.6	
	16	18.13	30.6217	68.901	19.5	7.55	20.6	13.7	
	17	20	30.3865	51.933	16.3	18.7	18.8	6.24	
	18	34.89	33.8271	3.046	32.7	6.23	40.0	14.8	
	19	65.85	34.7358	47.25	67.6	2.59	69.1	4.99	
	20	25.64	31.1359	21.435	26.2	1.98	20.5	19.9	
	21	15.6	30.3802	94.745	17.6	12.8	18.8	20.2	
	22	14.9	30.5754	105.204	15.7	5.62	19.0	27.4	
	23	32.68	32.3982	0.862	32.1	1.88	29.9	8.39	
	24	79.05	34.8936	55.859	79.1	0.000029	77.7	1.68	
	25	35.36	32.3736	8.446	33.1	6.30	30.3	14.4	
	26	34.37	32.5481	5.301	34.3	0.149	30.3	11.8	
	27	63.5	33.5591	47.151	63.0	0.871	44.4	30.1	
	28	11.12	30.8219	177.175	12.1	8.89	19.9	78.7	
	29	13.91	33.568	141.323	12.7	8.77	40.0	188	
	30	33.33	33.23	0.3	36.0	8.04	41.7	25.1	
	31	37.81	32.4452	14.189	37.1	1.93	37.8	0.0841	
	32	32.09	32.19	0.312	31.7	1.32	30.0	6.63	
	33	52.89	33.1232	37.373	52.9	0.0276	38.6	27.1	
	34	10.02	30.396	203.353	12.0	20.1	18.5	84.9	
	35	32.14	32.2773	0.427	31.0	3.49	30.1	6.26	
Prediction samples	36	39.04	32.2065	17.504	24.8	36.5	28.9	25.9	
	37	51.88	32.7885	36.799	44.3	14.7	39.3	24.2	
	38	26.6	31.6331	18.921	26.5	0.276	27.0	1.49	
	39	14.77	30.447	106.141	12.9	12.6	19.8	3.3.8	

Regression Problem for Calculating the Recovery Factor (RF)

Using the 35 learning samples with y^*= RF and by SVR, ANN and MRA, three functions of RF (y) with respect to 8 independent variables (x_1, x_2, …, x_8) have been constructed, corresponding to Eq. (A1), Eq. (A3) and Eq. (A5), respectively. Substituting the values of 8 independent variables given by the 35 learning samples and 4 prediction

sample (Table 1) in the three functions, respectively, the RF (y) of each sample is obtained (Table 3).

From Table 3 and based on the definition of result availability above, it can be seen that only ANN is applicable since its R(%) value is 5.89, whereas SVR and MRA are inapplicable because their R(%) values are 68.9 and 38.4, respectively (Seeing in Table 5).

Classification Problem for Determining the Recovery Factor Classification (RFC)

Using the 35 learning samples with y^*=RFC and by SVC, NBAY and BAYSD, three functions of RFC (y) with respect to 8 independent variables (x_1, x_2, …, x_8) have been constructed, corresponding to Eq. (A6), Eq. (A9) and Eq. (A11), respectively. Substituting the values of 8 independent variables given by the 35 learning samples and 4 prediction samples (Table 1) in the three functions, respectively, the RFC (y) of each sample is obtained (Table 4).

TABLE 4 PREDICTION RESULTS OF RECOVERY FACTOR CLASSIFICATION (RFC) IN 39 GLOBAL OILFIELDS

Sample type	Sample No.	RFC						
		y^*	Classification algorithm					
			SVC		NBAY		BAYSD	
			y	R(%)$_i$	y	R(%)$_i$	y	R(%)$_i$
Learning samples	1	5	5	0	5	0	5	0
	2	5	5	0	5	0	3	40
	3	3	3	0	5	66.7	5	66.7
	4	5	5	0	4	20	5	0
	5	1	1	0	1	0	3	200
	6	2	2	0	2	0	2	0
	7	3	3	0	3	0	3	0
	8	5	5	0	5	0	5	0
	9	5	5	0	5	0	5	0
	10	3	3	0	5	66.7	5	66.7
	11	2	2	0	2	0	2	0
	12	3	3	0	1	66.7	3	0
	13	3	3	0	3	0	3	0
	14	3	3	0	3	0	3	0
	15	4	4	0	4	0	4	0
	16	5	5	0	5	0	5	0
	17	5	5	0	5	0	5	0
	18	3	3	0	5	66.7	3	0
	19	1	1	0	1	0	1	0
	20	4	4	0	5	25	4	0
	21	5	5	0	5	0	5	0
	22	5	5	0	5	0	5	0
	23	3	3	0	5	66.7	5	66.7
	24	1	1	0	1	0	1	0
	25	3	3	0	3	0	3	0
	26	3	3	0	3	0	3	0
	27	1	1	0	5	400	3	200
	28	5	5	0	5	0	5	0
	29	5	5	0	5	0	3	40
	30	3	3	0	3	0	3	0
	31	3	3	0	3	0	3	0
	32	3	3	0	3	0	3	0
	33	1	1	0	1	0	3	200
	34	5	5	0	5	0	5	0
	35	3	3	0	3	0	3	0
Prediction samples	36	3	3	0	5	66.7	5	66.7
	37	1	1	0	3	200	5	400
	38	4	4	0	3	25	4	0
	39	5	5	0	5	0	5	0

From Table 4 and based on the definition of result availability above, it can be seen that only SVC is applicable since its R(%) value is 0, whereas NBAY and BAYSD are inapplicable because their R(%) values are 24.7 and 34.5, respectively (Seeing in Table 5).

Summary of the Case Study

Table 5 summarizes the applicability of each algorithm in the case study.

TABLE 5 SUMMARY OF THE CASE STUDY IN 39 GLOBAL OILFIELDS

Problem type	Algorithm	Mean absolute relative residual			Time consuming on PC[a]	Results availability
		R_1(%)	R_2(%)	R(%)		
Regression	SVR	71.6	44.8	68.9	3 s	Inapplicable
	ANN	4.74	1.60	5.89	30 s	Applicable
	MRA	40.4	21.4	38.4	<1 s	Inapplicable
Classification	SVC	0	0	0	5 s	Applicable
	NBAY	22.2	72.9	24.7	<1 s	Inapplicable
	BAYSD	61.2	55.4	34.5	1 s	Inapplicable

a. This program ran on a PC with configuration: HP Z230, Windows 7 (64 bit), Intel E3-1231, 3.40 GHz, 16GB RAM.

Conclusions

The purpose of this paper is to demonstrate how to select proper algorithms in three algorithms (SVR, ANN, MRA) for recovery factor regression and/or three algorithms (SVC, NBAY, BAYSD) for recovery factor classification, then to effectively predict it by appropriate DM algorithms and related data. From the aforementioned case study, four conclusions can be drawn as follows:

1) the proposed total mean absolute relative residual for all samples (R(%)) to express solution accuracy of an algorithm is practical;
2) the proposed result availability of a given algorithm application (applicable if R(%)<10, and inapplicable if R(%)≥10) is practical;
3) the preferable algorithm is ANN for recovery factor regression, while the preferable algorithm is SVC for recovery factor classification;
4) oil and gas recovery factor can be effectively predicted by appropriate DM algorithms, which has great significance for oil industry.

ACKNOWLEDGMENT

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

REFERENCES

[1] Altiparmak F, Dengiz B, and Bulgak AA. "Buffer allocation and performance modeling in asynchronous assembly system operations: An artificial neural network metamodeling approach." *Applied Soft Comput,* 7(3), 946–956, 2007.

[2] Boser BE, Guyon I, and Vapnik V. "A training algorithm for optimal margin classifiers." In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM Press, pp 144–152, 1992.

[3] Brown PJ, Kenward MG, and Bassett EE. "Bayesian discrimination with longitudinal data." *Biostatistics*, 2(4), 417–432, 2001.

[4] Chang C, and Lin C. "Training ν-support vector classifiers: Theory and algorithms." *Neural Comput*, 13(9), 2119–2147, 2001.

[5] Chang C, and Lin C. "Training ν-support vector regression: Theory and algorithms." *Neural Comput*, 14(8), 1959–1977, 2002.

[6] Chang C, and Lin C. "LIBSVM: a library for support vector machines, Version 3.1." Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2011.

[7] Chatterjee S, Hadi AS, and Price B. Regression Analysis by Examples, 3rd ed. New York, NY, USA. John Wiley & Sons Inc, 2000.

[8] Choi B, Lee JH, and Kim DH. "Solving local minima problem with large number of hidden nodes on two-layered feed-forward artificial neural networks." *Neurocomputing*, 71(16-18), 3640–3643, 2008.

[9] Cortes C, and Vapnik V. "Support-vector network." *Machine Learning*, 20(3), 273–297, 1995.

[10] Crisp DJ, and Burges CJC. "A geometric interpretation of ν-SVM classifiers." In: Solla S, Leen T, Müller KR (Eds), Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, V 12, pp 244–250, 2000.

[11] Cristianini N, and Shawe-Taylor J. "An Introduction to Support Vector Machines and other Kernel-based Learning Methods." Cambridge University Press, Cambridge, UK, 2000.

[12] Denison DGT, Holmes CC, Mallick BK, and Smith AFM. Bayesian Methods for Nonlinear Classification and Regression. Chichester, England, UK: John Wiley & Sons Inc, 2002.

[13] Ela MAE, Sayyouh H, and Tayeb ESE. "An integrated approach for the application of the enhanced oil recovery projects." *Journal of Petroleum Science Research*, 3(4), 176–188, 2014.

[14] Domingos P, and Pazzani M. "On the optimality of the simple Bayesian classifier under zero-one loss." *Machine Learning*, 29(2-3), 103–130, 1997.

[15] Güler İ, and Übeyli ED. "Detection of ophthalmic artery stenosis by least-mean squares backpropagation neural network." *Comput Biol Med*, 33(4), 333–343, 2003.

[16] Han J, Kamber M, and Pei J. Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.

[17] Hecht-Nielsen R. "Theory of the backpropagation neural network." In: Proceedings of the Int Joint Conf on Neural Networks, Washington, USA, pp 593–605, 1989.

[18] Hush D R, and Horne B G. "Progress in supervised neural networks." *IEEE Sig Proc Mag*, 10(1), 8–39, 1993.

[19] Lee JH, and Yang SH. "Statistical optimization and assessment of a thermal error model for CNC machine tools." *Int J Mach Tool Manufact*, 42(1), 147–155, 2002.

[20] Li D, and Shi G. "Data mining in petroleum upstream——the use of regression and classification algorithms." *Sci J Ear Sci*, 5(2), 33–40, 2015.

[21] Logan TP, and Gupta AK. "Bayesian discrimination using multiple observations." *Communications in Statistics-Theory and Methods*, 22(6), 1735–1754, 1993.

[22] Mi S, and Shi G. "The use of regression and classification algorithms for layer productivity prediction in naturally fractured reservoirs." *Journal of Petroleum Science Research*, 4(2), 65–78, 2015.

[23] Rumelhart DE, Hinton GE, and Williams RJ. "Learning internal representations by error propagation." In: Rumelhart DE, McClelland JL (Eds), Parallel Distributed Processing, MIT Press, Cambridge, MA, USA, Volume 1, pp 318–362, 1986.

[24] Schölkopf B, Smola AJ, Williamson RC, and Bartlett PL. "New support vector algorithms." *Neural Comput*, 12(5), 1207–1245, 2000.

[25] Salehi SM, and Honarvar B. "Automatic identification of formation lithology from Well Log Data: A Machine Learning Approach". *Journal of Petroleum Science Research*, 3(2), 73-82, 2014.

[26] 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.

[27] Shi G. Data Mining and Knowledge Discovery for Geoscientists. USA: Elsevier Inc, 2013.

[28] Shi G. "Optimal prediction in petroleum geology by regression and classification methods." *Sci J Inf Eng*, 5(2), 14–32, 2015a.

[29] Shi G. "Prediction of methane inclusion types using support vector machine." *Sci J Ear Sci*, 5(2), 18–27, 2015b.

[30] Shi G, Zhou X, Zhang G, Shi X, and Li H. "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." *Mar Petrol Geol*, 21(3), 411–420, 2004.

[31] Shi G, and Yang X. "Optimization and data mining for fracture prediction in geosciences." *Procedia Comput Sci*, 1(1), 1353–1360, 2010.

[32] Shi G, Zhu Y, Mi S, Ma J, and Wan J. "A big data mining in petroleum exploration and development." *Adv Petrol Expl Devel*, 7(2), 1–8, 2014.

[33] Shi G, Ma J, and Ba D. "Optimal selection of classification algorithms for well log interpretation." *Sci J Con Eng*, 5(3), 37–50, 2015.

[34] Singh J, Shaik B, Singh S, Agrawal VK, Khadikar PV, Deeb O, and 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.

[35] Tabach EE, Lancelot L, Shahrour I, and Najjar Y. "Use of artificial neural network simulation metamodelling to assess groundwater contamination in a road project." *Math Comput Model*, 45(7-8), 766–776, 2007.

[36] Tan P, Steinbach M, and Kumar V. Introduction to Data Mining. Boston, MA, USA: Pearson Education, 2005.

[37] Vapnik V. Statistical Learning Theory. New York, NY, USA: John Wiley Inc, 1998.

[38] Zhang Q, and Shi G. "Economic evaluation of waterflood using regression and classification algorithms." *Adv Pet Expl Devel*, 9(1), 1–8, 2015.

Dawei Li, Beijing, the People's Republic of China, is a senior engineer of PetroChina Research Institute of Petroleum Exploration & Development, Beijing, China. He holds bachelor, master and doctor degrees from China University of Geosciences, Wuhan and Beijing, respectively, all in Petroleum Geology. His research interests include geophysics, petroleum migration, basin modeling, petroleum resource evaluation and IT. During the past few years, he has spent much time on data collection, cleaning, database construction, application and mining in petroleum industry, having published several related papers on these domains. He has been working for more than 20 years in different departments of PetroChina (such as Bureau of Geophysical Prospecting, Research Institute of Petroleum Exploration and Development). He has published several books and more than 40 articles. Two of the major published books are "Tectonic types of Oil and Gas Basins in China" (Beijing: Petroleum Industry Press, 2003) and "Neotectonism and Hydrocarbon Accumulation in Huanghua Depression, China (in Chinese)" (Wuhan, Hubei Province: China University of Geosciences Press, 2006). Three of his important published articles are: "Ten critical factors for the evaluation of enterprise informatization benefits" (in Chinese, 2014), "Data mining in petroleum upstream—the use of regression and classification algorithms" (2015), and "Preprocessing of the data mining based on global oil and gas field database" (in Chinese, 2015). He ever attended or led the construction and application of several large-scale information systems for PetroChina, such as "E&P Technical Data Management System (A1)" and "Global Petroleum Resources Information System (GRIS)", etc. Dawei Li is a member of China Geological Society and China Petroleum Society.

Guangren Shi is a professor of PetroChina Research Institute of Petroleum Exploration & Development, Beijing, China.

Appendix A: Methods of the Six Algorithms in the Research

The following will discuss the six algorithms (SVR, ANN, MRA, SVC, NBAY, BAYSD). Because support vector machine (SVM) has both of classification (SVC) and of regression (SVR) algorithms, 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 ANN, though this problem in ANN is rare if ANN is properly coded (Shi, 2013). The SVM procedure was established in the 1990's, and included two principal algorithms: 1) SVC, such as the binary classification (e.g. Boser et al., 1992; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Shi, 2009; Shi and Yang 2010; Chang and Lin, 2011; Shi et al., 2014), and the ν-binary classification (Crisp and Burges, 2000; Schölkopf et al., 2000; Chang and Lin, 2001); and 2) SVR, such as the ε-regression (e.g. Vapnik, 1998; Chang and Lin, 2011; Salehi and Honarvar, 2014; Shi et al., 2014), and the ν-regression (e.g. Schölkopf et al., 2000; Chang and Lin, 2002). In the two case studies, the binary classification for SVC and the ε-regression for SVR 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 (Chang and Lin, 2011), and thus the kernel function used in SVC and SVR is the RBF.

Regression Algorithms

SVR, ANN and MRA use the same known parameters, and share the same unknown to be predicted. Only the real-type results calculated by each algorithm are different.

1) SVR

A technique of SVR, ε-regression (Chang and Lin, 2011; Shi, 2013), has been employed. The formula created by this technique is an expression with respect to a vector $\boldsymbol{x}$, which is so-called a nonlinear function $y=SVR(x_1, x_2, \ldots, x_m)$:

$$y = \sum_{i=1}^{n}\left[(\alpha_i^* - \alpha_i)\exp(-\gamma \parallel \boldsymbol{x} - \boldsymbol{x}_i \parallel^2)\right] + b \tag{A1}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ are the vector of Lagrange multipliers, $\boldsymbol{\alpha}=(\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\boldsymbol{\alpha}^*=(\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)$, $0\le\alpha_i\le C$ and $0\le\alpha_i^*\le C$ where C is the penalty factor, and the constraint $\sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0$; $\exp(-\gamma \parallel \boldsymbol{x} - \boldsymbol{x}_i \parallel^2)$ 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 $\boldsymbol{x}_i$. These free $\boldsymbol{x}_i$ are those vectors corresponding to $\alpha_i>0$ and $\alpha_i^*>0$, on which the final SVR model depends.

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

$$\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*}\left\{\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)y_i - \varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*) - \frac{1}{2}\cdot\sum_{i,j=1}^{n}\left[(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\exp(-\gamma\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)\right]\right\} \tag{A2}$$

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

It is noted that in the two case studies the formulas corresponding to Eq. A1 are not concretely written out due to their large size.

2) ANN

The ANN procedure has been widely applied since the 1980's (e.g. Rumelhart et al., 1986; Hecht-Nielsen, 1989; Güler and Übeyli, 2003; Shi et al., 2004; Altiparmak et al., 2007; Tabach et al., 2007; Choi et al., 2008; Shi, 2013), and the application of ANN is still predominant. The formula created by ANN is an expression with respect to m parameters ($x_1, x_2, \ldots, x_m$) (Shi, 2013):

$$y=ANN(x_1, x_2, \ldots, x_m) \tag{A3}$$

where ANN is a nonlinear function, which cannot be expressed as a usual mathematical formula and so is an implicit expression. ANN consists of one input layer, one or more hidden layers, and one output layer. In Case study 1, 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 $N_{hidden}=2(N_{input}+N_{output})-1$ 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.

The term back-propagation refers to the way (Güler and Übeyli, 2003), 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 ANN 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 (Hush and Horne, 1993) is

$$RMSE(\%)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2}\times 100 \tag{A4}$$

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 (e.g. Chatterjee et al., 2000; Lee and Yang, 2002; Shi et al., 2004; Singh et al., 2008; Shi, 2013), 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 ($x_1, x_2, \ldots, x_m$), plus a constant term, which is so-called a linear function $y=MRA(x_1, x_2, \ldots, x_m)$ (Shi, 2013):

$$y=b_0+b_1x_1+b_2x_2+\ldots+b_mx_m \tag{A5}$$

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. A5 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. A5 is solved via m iterations.

Classification Algorithms

SVC, NBAY and BAYSD use the same known parameters, and also share the same unknown to be predicted. Only the integer-type results calculated by each algorithm are different.

1) SVC

A technique of SVC, the binary classifier (Chang and Lin, 2011; Shi, 2013), has been employed. The formula created by this technique is an expression with respect to a vector $\boldsymbol{x}$, which is so-called a nonlinear function $y=SVC(x_1, x_2, \ldots, x_m)$:

$$y=\sum_{i=1}^{n}\left[y_i\alpha_i \exp(-\gamma \parallel \boldsymbol{x}-\boldsymbol{x}_i \parallel^2)\right]+b \tag{A6}$$

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 $\sum_{i=1}^{n} y_i\alpha_i=0$; $\exp(-\gamma \parallel \boldsymbol{x}-\boldsymbol{x}_i \parallel^2)$ 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 $\boldsymbol{x}_i$. These free $\boldsymbol{x}_i$ are those vectors corresponding to $\alpha_i>0$, on which the final SVC model depends.

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

$$\max_{\boldsymbol{\alpha}}\left\{\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\cdot\sum_{i,j=1}^{n}\left[\alpha_i\alpha_j y_i y_j \exp(-\gamma \parallel \boldsymbol{x}_i-\boldsymbol{x}_j \parallel^2)\right]\right\} \tag{A7}$$

It is noted that in the two case studies the formulas corresponding to Eq. A6 are not concretely written out due to their large size.

2) NBAY

The NBAY procedure has been widely applied since the 1990's, and widely applied in this century (e.g. Domingos and Pazzani, 1997; Shi, 2013). 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, …, x_m) (Tan et al., 2005; Han et al., 2012, Shi, 2013):

$$N_l(\boldsymbol{x})=\prod_{j=1}^{m}\left\{\frac{1}{\sigma_{jl}\sqrt{2\pi}}\exp\left(\frac{-(x_j-\mu_{jl})^2}{2\sigma_{jl}^2}\right)\right\}\quad (l=1,2,\ldots,L) \tag{A8}$$

where l is the class number, L is the number of classes, $N_l(\boldsymbol{x})$ is the discrimination function of the l^{th} class with respect to $\boldsymbol{x}$, 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. A8 is so-called a naïve Bayesian discrimination function.

Once Eq. A8 is created, any sample shown by Eq. 1 or Eq. 3 can be substituted in Eq. A8 to obtain L values: N_1, N_2, …, N_L. If $N_{l_b}=\max_{1\le l\le L}\{N_l\}$, then

$$y=l_b \tag{A9}$$

for this sample.

Eq. A9 is so-called a nonlinear function $y=NBAY(x_1, x_2, \ldots, x_m)$.

3) BAYSD

The BAYSD procedure has been widely applied since the 1990's, and widely applied in this century (e.g. Logan and Gupta, 1993; Brown et al., 2001; Denison et al., 2002; Shi, 2013). The following introduces a BAYSD technique, i.e. the successive Bayesian discrimination. The formula created by this technique is a set of nonlinear combinations with respect to m parameters (x_1, x_2, …, x_m), plus two constant terms (Shi, 2013):

$$B_l(\boldsymbol{x}) = \ln(p_l) + c_{0l} + \sum_{j=1}^{m} c_{jl} x_j \quad (l = 1, 2, ...L) \tag{A10}$$

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 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 l^{th} 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. Eq. A10 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. A10 is solved via m iterations.

Once Eq. A10 is created, any sample shown by Eq. 1 or Eq. 3 can be substituted in Eq. A10 to obtain L values: B_1, B_2, …, B_L. If $B_{l_b} = \max_{1 \le l \le L} \{B_l\}$, then

$$y = l_b \tag{A11}$$

for this sample.

Eq. A11 is so-called a nonlinear function $y = BAYSD(x_1, x_2, \ldots, x_m)$.

Among the aforementioned six algorithms, each of MRA and BAYSD has a special function: the results can be used as auxiliary data to show the dependence of the predicted value (y) on parameters (x_1, x_2, …, x_m), which can be used for dimensional reduction in DM (Shi and Yang, 2010; Shi, 2013). This dependence obtained by BAYSD is more reliable since BAYSD is nonlinear while MRA is linear.