The implementation of support vector machines for solving in oil wells

Zayar Aung
Applied Mathematics and Artificial Intelligence
National Research University
Moscow Power Engineering Institute (MPEI)
Moscow, Russia
zayaraung53@gmail.com

Ye Thu Aung
Applied Mathematics and Artificial Intelligence
National Research University
Moscow Power Engineering Institute (MPEI)
Moscow, Russia
yethuaungg55@gmail.com

Mihaylov Ilya Sergeevich
Applied Mathematics and Artificial Intelligence
National Research University
Moscow Power Engineering Institute (MPEI)
Moscow, Russia
fr82@mail.ru

Phyo Wai Linn
Dept. Applied Mathematics
Moscow State Technical University (Stankin)
Moscow, Russia
phyowailinnmipt@gmail.com

Abstract

The article deals with the problem of timely forecasting and classification of problems that arise in the process of well construction remains relevant. It is necessary to create a new methodology that should help drilling personnel to make timely decisions about possible problems in the drilling process on the basis of real-time data analysis, which will increase efficiency and reduce drilling costs accordingly.

Keywords*: drilling complications, machine learning, neural network, efficiency improvement, gradient boosting, classification*

I. INTRODUCTION

With the development of oil well digitization, both the data source for mass production parameters and the real-time data collection method support oil production with an optimized solution [1]. Using machine learning to improve, combine, modify, improve applications, and optimize oil well data analysis is a new smart scientific method of the oil well data analysis system. Currently, the parameters of an oil well used in the data analysis algorithm are relatively simple, in the absence of polyphyletic parameters, a standard for evaluating and terminating data [2]. In addition, in some oil wells that have entered the middle or later periods of the high water stage, features such as low permeability and resistivity of the complex accumulation layer may cause the General manual analysis and linear analysis to be invalid [3]. From the perspective of intelligent machine learning, a nonlinear SVM classification algorithm is proposed in this paper, the structure of the data development system and the pattern recognition model for polyphyletic parameters are constructed, and the use of SVM through a high-dimensional spatial feature map and hyperplane optimized classification allows solving the problem of analysing nonlinear parameters of oil wells and pattern recognition.

II. PATTERN RECOGNITION OF OIL WELLS

In the course of oil production, the monitoring centre collects, transmits, analyses and provides real-time data on the flow rate of oil and gas for oil production, product watering, pressure, temperature, electrical voltage, electric current and load, as well as other primary parameters, which helps the administrator understand the operating conditions of the oil well and ensure its operation in a high-efficiency and low operating flow mode [4]. As a rule, these parameters also include peak values of electric current and voltage, pump pressure, back pressure, oil pressure and pressure in the annular space of the well.

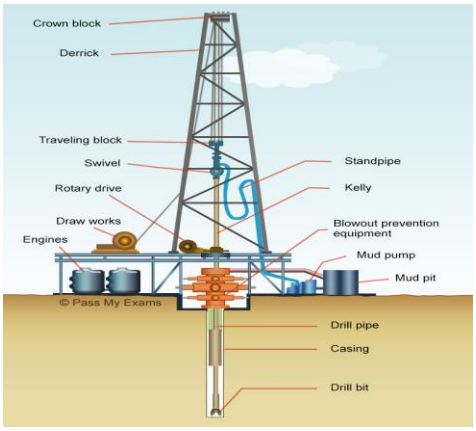

Figure 1. Intelligent systems in oil fields

This data is transmitted to the automated control system in real time. After performing a linear approximation and forecasting of the obtained data, the decision-maker can assess the state of the well at the moment and predict its behaviour in the future, and take appropriate compensating control actions.

III. NONLINEAR SVM

The kernel method allows for solving the problem of nonlinear classification using a nonlinear transformation [7]. Provided that the input space is Euclidean and the feature space is Hilbert, the kernel method means the product of feature vectors obtained by converting input data from the input space to the feature space. Using the kernel method to study nonlinear data in order to obtain a nonlinear SVM. The entire procedure is the operation of the linear SVM method in a multidimensional feature space.

The General idea is to use a nonlinear transformation to change the input space into a feature space, which can transform the hypersurface model in the source space into a hyperplane in the feature space. This means that a nonlinear classification problem in the original space is transformed into a problem that can be solved b y a linear SVM in the feature space [5].

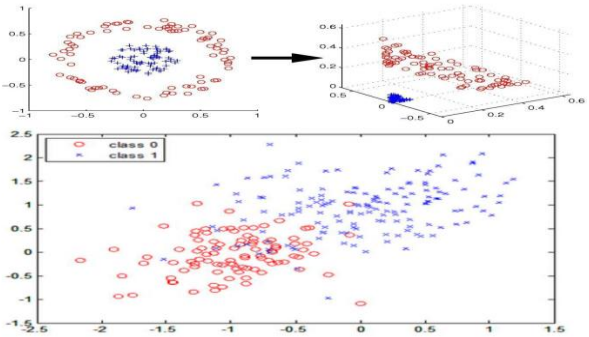

Figure 2. Using the kernel method to solve a nonlinear problem

IV. SUPPORT VECTOR MACHINE

The SVM model builds a hyperplane or set of hyperplanes in a multidimensional space called the feature space, which can be used for classification or regression. Its advantages over other machine learning methods include greater generalization capability, strong noise immunity, and less learning time (Vapnik V. 1995; Ani-fowose and Abdulraheem 2011). The SVM approach was developed in 1992 by a company Vanpik in collaboration with the Laboratory of Bell Laboratories. The SVM model is a set of interrelated managed learning methods used for classification and regression [6]. The SVM principle is based on statistical learning theory and structural minimization, which has shown better performance than the usual empirical risk minimization used by many machine learning methods (James Lara 2008).

$$min_{w,b,\varepsilon} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\varepsilon_i. \quad (1)$$

$$\text{s.t. } y_i(wx_i + b) \geq 1 - \varepsilon_i \quad (2)$$

Where C is the penalty parameter. Increasing C also increases the penalty for classification errors. You must adjust the target function to minimize the number of singular points while maximizing the offset from the hyper plane.

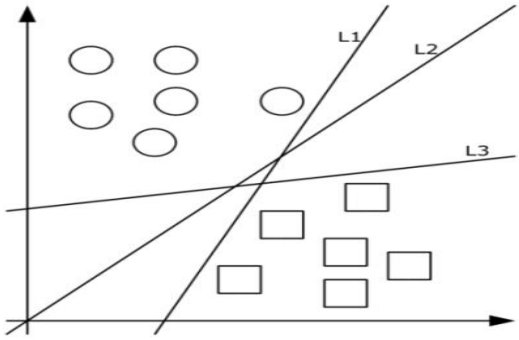

Figure 3. Hyperplanes and Support Vectors

V. LINEAR LOGISTIC REGRESSION GORITHM

The linear logistic regression algorithm is a classic classification method for studying statistics related to the linear logarithmic model [8]. This classification model is a conditional probability distribution P (Y / X), which is a judgment model. It can be obtained from the linear regression model hw (x)= $_wT_X$ and the sigmoid curve:

$$P(Y = 1|X) = \frac{1}{1+e^{-wx}}. \quad (3)$$

Where X is the input, Y is the output, W is the weighted coefficient, and WX is the internal product. The logistic regression distribution function and the density function are shown in Fig. 3. Logistic regression compares the difference between two conditional probabilities and classifies the training example x

into a large probability group. For the data of the training set it is possible to use maximum likelihood to estimate the parameters of the model to obtain the logistic model. The following assumptions are introduced [9].

$$P(Y = 1x) = f(x), P(Y\,0 = x)\,1 = f(x) \quad (4)$$

Likelihood function

$$\prod_{i=1}^{N} [f(x_i)]^{y_i} [1 - f(x)]^{1-y_i}. \quad (5)$$

Logarithm likelihood function

$$L(w) = \sum_{i=1}^{N} [y_i \log f(x_i) + (1 - y_i) \log(1 - f(x_i))]. \quad (6)$$

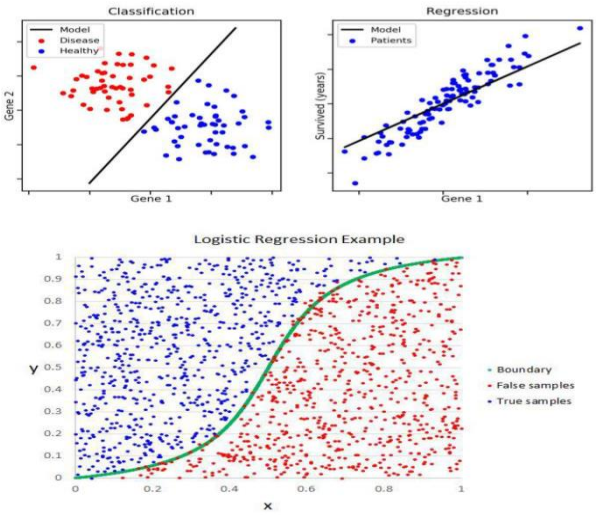

Figure 4. Logistic regression distribution function and density function

VI. IMPLEMENTATION AND RESULTS OF THE EXPERIMENT

The main purpose of the study is to evaluate the effectiveness of oil well planning. The efficiency of the system is the most important factor in the quality of the production system. The efficiency of a production system is the ratio of the useful amount of liquid produced to the power consumed per unit of time, which is a significant factor in production. As a result of the experiment, the efficiency of the system was chosen as the target factor. It is assumed that the system efficiency value above 45% is positive, less than 45% is negative.

In data mining, parameters such as pump load, temperature, and electrical voltage are suitable for solving the classification problem in the evaluation model. When analyzing the efficiency of the pumping system, the factors that affect it listed in the table are taken into account.1 and table. 2 [10].

Table 1. Oil well parameters

Parameter	Unit	Parameter	Unit
Reactive power	[KW]	Depth	[m]
Oil pressure	[MPa]	Period of work	[/]
Max pressure	[MPa]	Max load	[KN]
Min pressure	[MPa]	Min load	[KN]
Power factor	[1]	Production pressure	[MPa]
Voltage	[V]	Active power	[KW]
Current	[A]	Max active power	[KW]

Table 2. Oil well production parameters

Parameter	Unit	Parameter	Unit
Doppler velocity (array)	[Hz]	Liquid Consumption	[m^3/day]
Gas void fraction (array)	[%]	Gas Consumption	[m^3/day]
Sound velocity	[m/s]	Water cut	[%]
Fluid Pressure	[MPa]	Temperature	[℃]

Primary data were obtained for oil fields in the Perm region, for oil wells and booster pumping stations for a long period of their operation [11].

The first block of data presented in table 1 was relatively easy to obtain, since these parameters are measured directly by the corresponding sensors. The data shown in table 2 are the results of measurements of an innovative ultrasonic multiphase flow meter. It was installed in oil wells and booster pumping stations and consists of a vertical measuring tube with two different calibrated sections, four Doppler sensors, four gas void fraction (GWF) sensors, two sound velocity sensors, a thermometer, a sensor, and a computing unit with a multi-layer mathematical model. This model sets the boundary between the primary data (Doppler speed, GWF, speed of sound, temperature, pressure) and the calculated data (liquid flow, gas flow, and water flow).

The main parameters of oil production efficiency are the flow rate of liquid, gas and water content. However, using the primary parameters, you can determine the flow mode of the mixture. There are four main types of flow modes: bubble, mucus, dispersed ring, and dispersed [12]. The flow mode shows the stability of the oil well operation, and it should also be taken into account when evaluating the efficiency [13].

To reduce the feature space, the integral values of the Doppler velocity and GWF can be calculated as the arithmetic mean of the four corresponding parameters. Parameters are measured at four points of two different calibrated cross sections of the pipe: in the center of the small cross section, on the periphery of the small cross section, in the center of the large cross section, on the periphery of the large cross section [14].

According to the mathematical model of the flow meter, the flow rate depends on the primary data of the twelve parameters. However, as an example in Fig.1 shows the relationship between the fluid flow rate and the integral Doppler velocity.

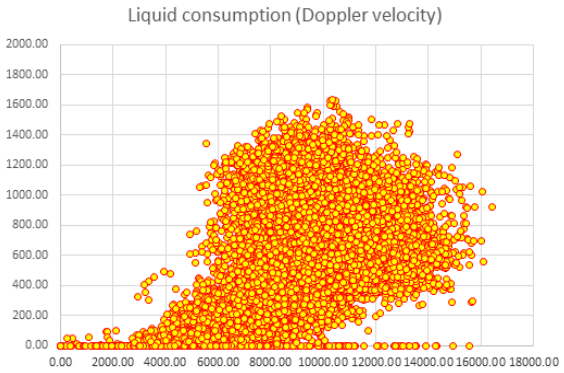

Figure 4. Dependence between liquid consumption and integral Doppler velocity

Figure 4 shows that when the Doppler velocity increases, the fluid flow increases. However, fluid flow also correlates with other parameters such as GVF, water flow, temperature, pressure, and others. And the main correlation between the flow rate of the liquid and the integral Doppler velocity is blurred under the influence of these parameters [15].

For rice. 2 shows the relationship between the gas flow rate and the integral fraction of gas voids.

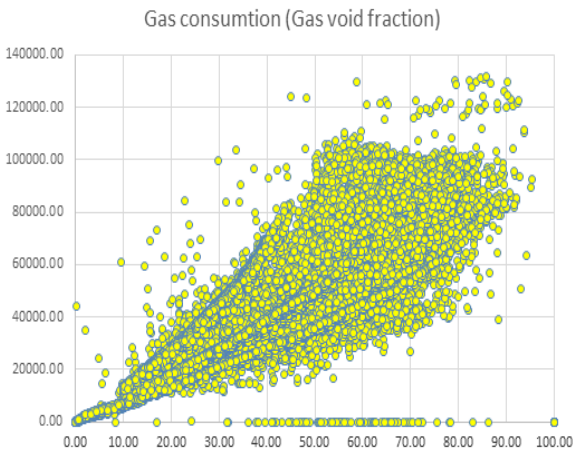

Figure 5. Dependence between gas consumption and integral gas void fraction

It follows from the figure 5, it is shown that the higher the GWF, the greater the gas consumption. However, gas consumption also depends on other parameters, such as fluid velocity, water content, temperature, pressure, and so on. the main correlation between gas consumption and GWF is also blurred by these parameters. Moreover, the greater the absolute value of the GWF, the greater the variation in gas flow [17].

Therefore, after analyzing cross-correlations in the training set, it was determined that all parameters should be used in the study.

To improve the results of the work performed in accordance with the received data, the following actions were performed.

1) to improve the efficiency of data collection, all possible related information was collected.
2) created an evaluation model for solving real problems.
3) an optimal modular scheme has been developed.
4) results are compared with real data and the model is updated.
5) the data set was pre-processed using anti-aliasing, normalization, and noise reduction techniques.
6) the evaluation of the obtained models was carried out.

VII. Classification Results

As a training set, 2019 examples of data from oil wells and pumping stations were selected, and 20 examples were selected as a test sample. According to experience, the penalty parameter C was set to 0.8, the RBF estimation function and the standard deviation 0.5 for the SVM model; the penalty parameter C=1 for LR. A comparison of projected and actual

performance indicators is shown in table 3. The experiment was performed in python on the example of oil well data using SVM and LR algorithms. Five years of measurement experience and primary and calculated data were obtained, where they were systematized and cleared.

№	Real values	Forecast LR	Forecast SVM	№	Real values	Forecast LR	Forecast SVM
1	0	1	0	11	0	0	1
2	0	0	0	12	0	0	0
3	0	0	0	13	0	0	0
4	0	0	0	14	0	0	0
5	0	0	0	15	0	0	0
6	0	1	1	16	1	1	1
7	1	0	1	17	0	1	0
8	1	1	1	18	0	0	0
9	0	1	1	19	0	0	0
10	1	1	1	20	1	1	1

Using the logistics model, 18 correct classifications were found with 90% accuracy that meet the forecast conditions. Using the PCA dimension reduction method, the data dimension was reduced from 17 to 2. The deviation of the result from the graph of real values is shown in Fig.6. 15 correct classification results were found, which means that the accuracy reaches 75%. Within the SVM model.

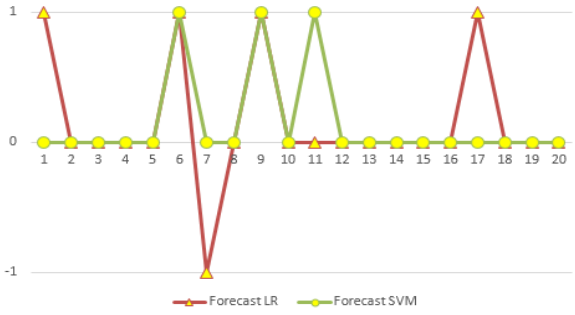

Figure 6. Results of the SVM and LR experiment

In the subject area of oil wells, data distribution is complicated by the high dimension of the information space, which can have a large impact on the collection of primary data. In this situation, there may be errors in the collection of one or more data types, as well as uneven distribution of data. Classic manual analysis, such as using charts, linear analysis, or logistic regression, does not provide qualitative classification. In this case, a support vector machine using the kernel method is better suited for nonlinear complex data processing. In this case, the current operating mode of oil wells is considered automatically based on a set of primary data that leads to lower or higher values of liquid and gas flow rates and to lower or higher efficiency values, without direct classification of oil wells by this parameter. However, it would be very useful to develop such a classification for predicting the behavior of oil wells and preventing accidents. For figure 6 deviation ratio between the real value and the forecast of LR is shown by triangles, and the deviation ratio between the real value and prediction of SVM are shown by circles.

VIII. Conclusion

The article presents a theoretical analysis of the support vector machine and logistic regression. It is shown that the nonlinear SVM algorithm works better than the linear LR algorithm when analyzing the oil well system and predicting its efficiency. In future studies based on the current review, it is necessary to develop a method of multiple classifications based on the support vector machine, which allows you to classify the original data set into several classes with the ability to assess the degree of proximity to each of these classes. For an oil field process service, it is very important to determine the multiphase flow mode (oil-water-gas), since, for example, even a high flow rate in an inappropriate mode can lead to a pump failure and an emergency stop of the oil well. However, if this situation could be classified at an early stage, it would avoid an emergency situation and thus preserve the efficiency of oil well operation.

REFERENCES

[1] Bashmakov, A. I., Bashmakov, I. A. Intellectual information technologies. Benefit // –M.: Publishing MGTU im. N.Uh. Bauman, 2005.

[2] Вапник В. Н., Статистическая теория обучения: Нью- Йорк: John Wiley \ & Sons, 1998, 740 С. RRDtool. URL:

http://oss.oetiker.ch/rrdtool/ (дата обращения: 25.04.2013).

[3] Kaufman L., Rousseeuw P. J., Нахождение групп в данных введение в кластерный анализ: NJ, Hoboken, USA: John Wiley \ & Sons, 2005, 355 p.

[4] Scholkopf Б., С. Платт Дж., Shawe-Тейлор Дж., Смола А.Дж., Уильямсон К. К., Нейронные вычисления, 2001, том. 13,С. 1443-1471.

[5] Лин ХС.- Ти, Линч.- J., Weng R. C., Машинное обучение,2007, Vol. 68, PP.

[6] Санчес-Фернандес М., Арен-Гарсия Дж., Перес-Крус Ф.,Сделки IEEE по обработке сигналов, 2004, вып. XX, нет. V, PP.

[7] Черкасских В., М. GPE фирмы Intel 2415, Спрингер-Верлаг, Берлин-Хайдельберг, 2002, с. 687-693.

[8] Ма Я., С. Перкинс, Тез. Доклад на 9-м симпозиуме ACM'03, Вашингтон, округ Колумбия, США, 2003, с. 613-618.

[9] Yeon Su Kim. Evaluation of the effectiveness of classification methods: comparative modeling. Expert systems with applications, 2009, 373 p.

[10] Hanuman Tota, Raghava Miriyala, Shiva Prasad Akula, K. Mrityunjaya RAO, Chandra Sekhar Vellanki, etc.. Performance comparison in classification algorithms using real data sets. Journal of computer science and systems biology, 2009, 02-01.

[11] honglin AO, Junsheng Cheng, Yu Yan, Dong HAK Chu-Ong. The method for optimizing the parameters of support vectors is based on the algorithm for optimizing artificial chemical reactions and its application for diagnosing roller bearing failures. Journal of vibration and control.2015 (12).

[12] Agrawal Rimjhim, Tukaram Of Dharbanga. Identification of the fault location in distribution networks using multi-class support vector machines. International journal of developing electric power systems.2012 (3).

[13] A. Snehal Mulay, P. R. Devale, G. V. Garje. Intrusion protection system using a support vector machine and a decision tree. International journal of computer applications.2010 (3).

[14] Wang Lijun, Lai Huicheng, Zhang Taiyi. Improved least squares support vector machine algorithm. Journal of information technology.2008 (2).

[15] R. Cogdill And P. Dardenne. Least squares support vectors for Chemometry: an introduction to andevaluation. Journal of near-infrared spectroscopy.2004 (2).

[16] Ke Lin, Anirban Basudhar, Sami Missum. Parallel construction of explicit boundaries using support vector machines. engineering calculation.2013 (1).

[17] Ashkan Musavian, Hojat Ahmadi, Babak, Sahai, Reza Labbafi. Support vector machine and K-nearest neighbor for unbalanced fault detection. Journal of quality in the field of maintenance.2014 (1).