Machine Learning Algorithm For Efficiency Management Of Oil Well

Shi-qi Bao[1, a], Zhi-jie Ding[2,b] ,Yun-yun Wu[3,c] ,Yue-ting Shi[4,d]

[1,2,3,4] Beijing Institute of Technology , Beijing ,China

[a]baoshiqi27@163.com, [b] kaputt@bit.edu.cn

Keywords: Machine learning ; pattern recognition ; computer classification; application of oil field

Abstract.

Based on machine learning technique and oil well efficiency project practical problem, to the complicated circumstance of oil well efficiency, non-linear machine learning support vector machines (SVM) shows a better analysis results than the classified prediction result of linear machine learning logistics regression (LR). This paper analyzed and derived the theorems and classification reason of logistics regression and support vector machines. The experiments calculated and compared the accuracies of these two algorithms under the same conditions, the result conforms the conclusion.

Introduction

With the digitalization development of oil well, both data source of mass production parameters and real time data collection technique support oil well production with optimized decision [1]. Using machine learning to clear up, integrate, convert, develop applications and optimize analysis of oil well data is a new reasonable scientific method of oil well data analysis system. Nowadays, the oil well parameters used in data analysis algorithm are relatively simple, in lack of polyphyletic parameters, evaluation standard and data redundancy [2]. Moreover, with some oil wells entered middle or later periods of high water cut stage, the features like low permeability and resistivity of complicated accumulation layer can cause general manual analysis and linear analysis invalid[3]. In the angle of intelligent machine learning, we propose a nonlinear SVM classification algorithm in this paper, building up the structure of data development system and pattern recognition model of polyphyletic parameters, using SVM through high-dimensional feature space map and hyperplane optimized classification can solve oil well nonlinear parameters analysis and pattern recognition issue.

1. Polyphyletic parameters oil well pattern recognition model

In the production of oil well, surveillance center collects, transfers, analyzes and releases the real time data of pressure, temperature, electric voltage, electric current and load, which helps the administrator to know the real time working conditions well and makes the oil well operate under high efficiency and low consumption condition [4-5]. For working conditions, they are mainly indicator diagrams, electric current peak value, equilibrium ratio, electric voltage, pump stroke number, pump pressure, return pressure, oil pressure and casing pressure. For production parameters, they are mainly daily fluid production and return oil temperature. These data are the source of the system which satisfies ordering, continuity and real-time ability. After linear data development of production parameter data source, the decision maker can get the technical qualification like system efficiency and working conditions at presentation layer.

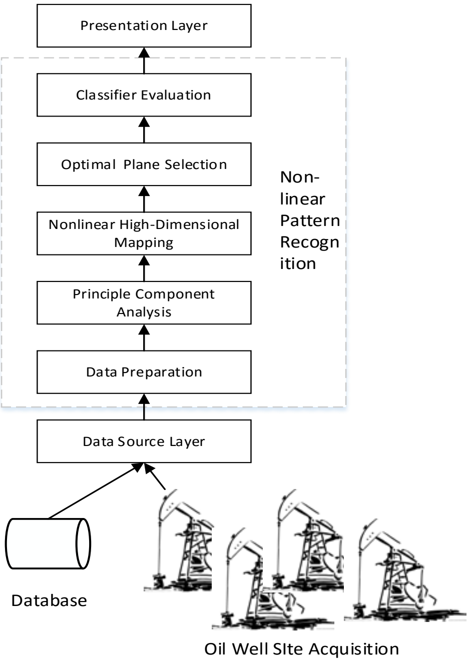
Presentation Layer
Classifier Evaluation
Optimal Plane Selection
Nonlinear High-Dimensional Mapping
Principle Component Analysis
Data Preparation
Non-linear Pattern Recognition
Data Source Layer
Database
Oil Well SIte Acquisition

Figure 1. Multiple parameters of oil well pattern recognition model [5]

2. Nonlinear SVM

2.1 Kernel method

Kernel method can solve the problem of nonlinear classification through nonlinear transform [6]. When the input space is Euclid space or disjoint set feature space is Hilbert space, kernel method means the inner product of feature vectors derived from the process which converts input data from input space to feature space. The study of nonlinear data can be done through this kernel method, and finally obtain the nonlinear SVM. The entire procedure equals to the implicit study of linear SVM in high-dimensional feature space.

Kernel method is shown in Figure 2. The general idea is using a nonlinear transform to change the input space into a feature space, which can convert the hypersurface model in original space into a hyperplane model in feature space. That means the nonlinear classification problem in original space is converted to a problem can be solved by linear SVM in feature space.

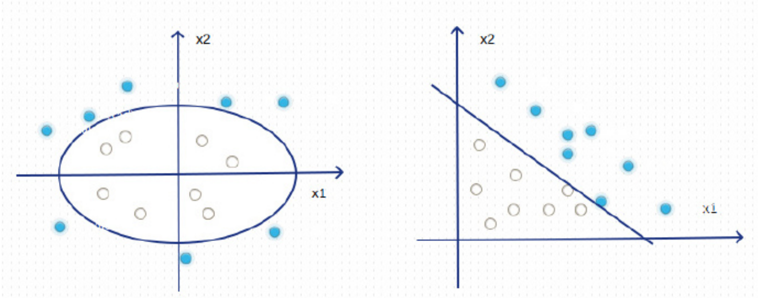

Figure 2. Using kernel method to solve nonlinear problem

2.2 SVM

The general idea of SVM is to solve the problem of correct classification of data set and maximize the geometric margin. There can be multiple separating hyperplanes but there are only one separating hyperplane with the maximum geometric margin. The direct explanation of maximizing the geometric margin is that the hyperplane with maximum geometric margin gained from classification equals to classify training data by sufficient certainty factor. Not only to classify correctly, but also separate the nearest points with sufficient certainty factor. This process can provide determined data with a good predictive ability, which is called generalization ability.

When coping with nonlinear problem, after converting into high-dimensional space, it is usually hard to find a hyperplane which can completely separate the data points, which means there are some singular points. But after wiping off those singular points, most of the left part are linearly separable. In order to solve this problem, we import a slack variable to the training sample. In the situation of soft margin, the SVM training problem can be:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \cdot \tag{1}$$

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

Where C is penalty parameter. When C increases, the mistake classification penalty increases as well. Adjust the target function in order to minimize the number of singular points meanwhile maximize the margin.

3. Linear logistic regression algorithm

Linear logistic regression algorithm is a classic classification method in statistics study which belongs to linear log model. It is a classification model represents by conditional probability distribution P(Y/X), which is a judgment model. It can be derived from the linear regression model $h_w(x)=w^Tx$ and the sigmoid curve:

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

Where x is the input, y is the output, w is weighted factor and wx is the inner product.

Logistic regression distribution function and density function is shown in Figure 3.

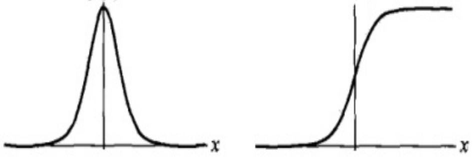

Figure 3. Logistic regression distribution function and density function

Logistic regression compares the difference between two conditional probabilities and sort the example x into the lager probability group.

To the provided training data set, we can use the maximum likelihood to estimate the model parameter to obtain the logistic model. Assuming

$$P(Y=1\,|\,x)=f(x),P(Y=0\,|\,x)=1-f(x)\,. \tag{4}$$

The likelihood function is

$$\prod_{I=1}^{N}[f(x_i)]^{yi}[1-f(x)]^{1-yi}\,. \tag{5}$$

The logarithm likelihood function

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

4. Experiment design and results

4.1 oil well system efficiency experiment design

System efficiency is the most important factor to qualify pump systems. Pump system efficiency is the ratio of useful power of lifting liquid to input power in a unit time, which is an essential factor of production. As a result, the experiment set system efficiency as the target factor. Assume when system efficiency is larger than 45% as positive, while smaller than 45% as negative.

In data mining, the parameters, such as displacement, load, temperature and voltage, of pump in oil exploration correspond to the classification mission in estimation model. With the analysis of pump's system efficiency, we can know that there can be influence factors listed in Table 1. The data in Table 1 was obtained from every oil well at the same time with a sample of 2000.

Table 1. Oil well parameters

Parameters	Unit	Parameters	Unit
Stroke	[m]	Mean reactive power	[kw]
Times of stroke	[/]	Oil pressure	[mpa]
Max load	[kn]	Max pressure	[mpa]
Min load	[kn]	Min pressure	[mpa]
Mean power factor	[/]	Return pressure	[mpa]
Mean active power	[kw]	Voltage	[v]
Max active power	[kw]	Current	[A]

According to the data and application development, the followings are the general procedure:

1 To guarantee the effectivity of data, we need to collect all possible related information according to the predicted oil well.
2 Pre-process the data with smoothing, normalization and denoising methods.
3 Build up a evaluation model to solve real problem.
4 Evaluate the features received.
5 Conform the modeling scheme.
6 Operate the experiment and compare the result with real oil well to upgrade the model.

4.2 classification results

Taking oil wells in Dagang Oil Field as example, the experiment operated in python by using SVM and LR algorithm. Taking 1980 oil wells as training data and the rest 20 oil wells as prediction. According to experience, the design parameter C=0.8, kernel function is RBF and standard deviation is 0.5 for SVM; the design parameter C=1…….for LR. The comparisons of predicted efficiency and real efficiency are listed in Table 2.

Table 2. Results of classification

number	real	LR predict	SVM predict	number	real	LR predict	SVM predict
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

4.3 result analysis

Under the logistic model, there are 15 correct classifications which means the accuracy reaches 75%. Under the SVM model, there are 18 correct classifications with the accuracy of 90%, which meets the prediction need. With the PCA dimensionality reduction method, we can reduce the 17 dimensional data to 2 dimensional in consideration of visualization, whose result shows in Figure 4. The set of whole points in Figure 4 means the determined dataset. Squares mean the correct classifications of SVM while stars mean that of LR. The overlapping parts are correct in both algorithms and red crosses are classification errors.

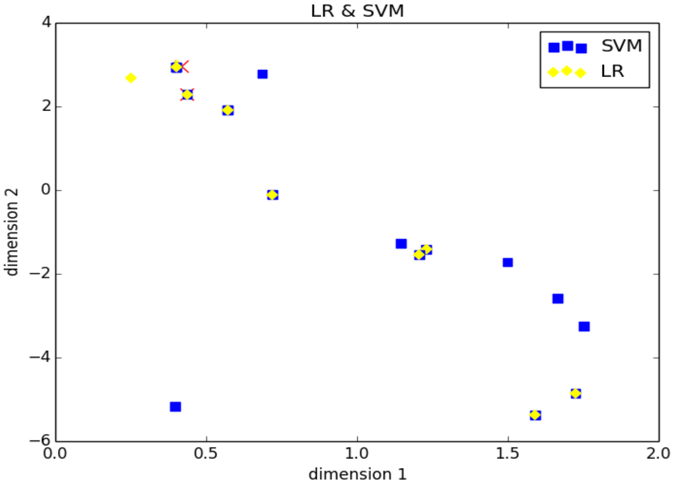

Figure 4. The experiment results of SVM and LR

Data distribution is more complicated to an oil well system because of the high dimensional and there can be a huge influence on data collection according to the environment. In this situation, the collection error of one or some kinds of data is possible, as well as the uneven data distribution. Classic manual analysis like indicator diagram and linear analysis like LR does not work well. Under this circumstance, maximized interval of sample points and SVM using kernel method are better for nonlinear complex data process.

5. Further work

The paper starts with theory algorithm analysis corresponding with project experiment result, in order to prove the nonlinear SVM algorithm works better than linear LR algorithm on oil well system analysis and efficiency prediction. As a result, when analyzing a complicated oil well system, nonlinear SVM may be a better choice. Due to the limitation of dataset, the experiments can be operated on one machine. But when the dataset reaches a level, then the processing method need to be changed into distribution method, which means we need distribution machine learning to improve it.

References

[1]Yong Soo Kim. Performance evaluation for classification methods: A comparative simulation study[J]. Expert Systems With Applications, 2009,373.

[2]Hanuman Thota, Raghava Naidu Miriyala, Siva Prasad Akula, K. Mrithyunjaya Rao, Chandra Sekhar Vellanki, et al.. Performance Comparative in Classification Algorithms Using Real Datasets[J]. Journal of Computer Science &amp; Systems Biology, 2009, 0201.

[3] HungLinh Ao, Junsheng Cheng, Yu Yang, Tung Khac Truong. The support vector machine parameter optimization method based on artificial chemical reaction optimization algorithm and its application to roller bearing fault diagnosis. Journal of Vibration and Control.2015(12).

[4] Rimjhim Agrawal, Thukaram Dhadbanjan. Identification of Fault Location in Distribution Networks Using Multi Class Support Vector Machines. International Journal of Emerging Electric Power Systems.2012(3).

[5] Snehal A. Mulay, P.R. Devale, G.V. Garje. Intrusion Detection System Using Support Vector Machine and Decision Tree. International Journal of Computer Applications.2010(3).

[6] Wang Liejun, Lai Huicheng, Zhang Taiyi. An Improved Algorithm on Least Squares Support Vector Machines. Information Technology Journal.2008(2).

[7] R. Cogdill, P. Dardenne. Least-squares support vector machines for chemometrics: an introduction andevaluation. Journal of Near Infrared Spectroscopy.2004(2).

[8] Ke Lin, Anirban Basudhar, Samy Missoum. Parallel construction of explicit boundaries using support vector machines. Engineering Computations.2013(1).

[9] Ashkan Moosavian, Hojat Ahmadi, Babak Sakhaei, Reza Labbafi. Support vector machine and K-nearest neighbour for unbalanced fault detection. Journal of Quality in Maintenance Engineering.2014(1).

[10] Long Zhang, Jianhua Wang. Optimizing parameters of support vector machines using team-search-based particle swarm optimization. Engineering Computations.2015(5).