Recognition of Oil Shale Based on LIBSVM Optimized by Modified Genetic Algorithm

Qihua Hu[1,*], Cong Wang[2], Xin Zhang[1] and Jingjing Fan[1]

[1]College of Geoscience and Surveying Engineering, China University of Mining & Technology, Beijing 100083, China; [2]Beijing Geotechnical Engineer Institute, Beijing, 10083, China

Abstract: In order to improved the speed, accuracy and generalization of oil shale recognition model with log dada, considering parameters of traditional SVM were chosen by experience, a LIBSVM recognition model with optimized parameters was proposed based genetic algorithm. First of all, all the samples data were processed to double type as LIBSVM tool needing, and the best normalization way was chosen through comparing different accuracies of various normalization ways. Secondly, the fitness value was calculated by the traditional LIBSVM. Finally, parameters C and g were optimized by genetic algorithm according the fitness value. The optimized LIBSVM oil shale recognition model was applied in northern Qaidam basin to identify oil shale, the results show that optimized recognition model is a tool of better generalization ability and the recognition accuracy reaches as much as 97.2806%. According to the popularization effects in the well area of same geology background, this optimized LIBSVM model is the best for now.

Keywords: Genetic algorithm, LIBSVM, log interpretation, oil shale recognition.

1. INTRODUCTION

As logging curves of oil shale indicates the characteristics of high resistivity, high interval transit time, high natural gamma, high neutron porosity and low volume density, the methods of recognition of oil shale are currently △LogR overlap method, combination method of multiple logging curves, and so on. However, applying these methods needs a large amount of complicated calculation, which are easily to make mistakes and low efficient. According to Zhu Jianwei's research in the application of identification of oil shale with logging curves, recognition and quantitative analysis of oil shale were both based on the linear relation between log response and actual stratum, but for the complexity of geology condition and sedimentary environment, adding with formation heterogeneity, there must be a nonlinear relation between them. According to Zhang Jiajia's research in recognition of oil shale with seismic method, although it had shown a better result, it was also difficult to use this method when there were no seismic data in some old area only with log data. Therefore, combined with other methods, a method based on a Library for Support Vector Machine (LIBSVM) modified by genetic algorithm was promoted to identify oil shale. Support Vector Machine (SVM) firstly proposed by Vapnik was a new machine learning method derived from principle of structure risk minimization and the VC dimension theory in statistical learning theory, and this method can be used in pattern classification and nonlinear regression with the global searching ability of genetic algorithm in optimization of complex system, a preferences method of LIBSVM was proposed to apply in recognition of oil shale in the northern Qaidam basin, and it offered a feasible and efficient method to apply in the recognition of oil shale in the same geology condition background.

There are many kinds of SVM toolboxes, and LIBSVM is currently the best toolbox of them, which was developed by professor Chi-Jen Lin from Taiwan University, It was designed as a package for pattern recognition and regression in SVM easily, simply and quickly. It not only offer a compiled executable file in Windows series system but also source code, which can be modified and updated conveniently in the application in other systems.

The classification principle of SVM is to finding an optimal hyper plane as a decision surface under a linearly separable condition, this decision surface will maximization dividing edges between positive examples and negative examples, thus it will achieve a best classification result under the lowest percent of misjudge examples. Assuming that training samples is (x_i, y_i) , i=1,2,…,N, and xi is input value, yi is expecting output, $y_i=\pm1$ separately expresses positive examples and negative examples, the formula indicating the decision surface is shown as below:

$$\omega^T x + b = 0 \qquad (1)$$

Formula (1) satisfy linearly separable samples (xi, yi) with :

$$y_i\left(\omega^T x + b\right) - 1 \geq 0$$

In the formula, x is input vector, ω is adjustable weight vector, b is bias. For the given weight vector and bias b, the interval space between the hyper plane defined by the formula (1) and the nearest data, ρ stands for it. xi is the support vector on the dividing edge, whose number is limited, and they can display the hyper plane. According to the principle of SVM, maximization of dividing edge between positive and negative examples is finding the max value of ρ. According to Li Yang's research in LIBSVM, x_1, x_2 separately stands for positive examples and negative examples, the interval space between them can be displayed as below:

$$dis=\frac{\omega}{||\omega||}\cdot\left(x_1-x_2\right)=\frac{2}{||\omega||} \qquad (2)$$

According to the formula (2), maximization of interval space turns to maximize $\frac{2}{P\omega P}$, also minimize $\frac{P\omega P}{2}$, turning it into minimize $\frac{P\omega P^2}{2}$ in order to calculate conveniently. As to find a hyper plane, which is actually to find ω and b satisfied minimization of $\frac{P\omega P^2}{2}$ confined with formula (1).

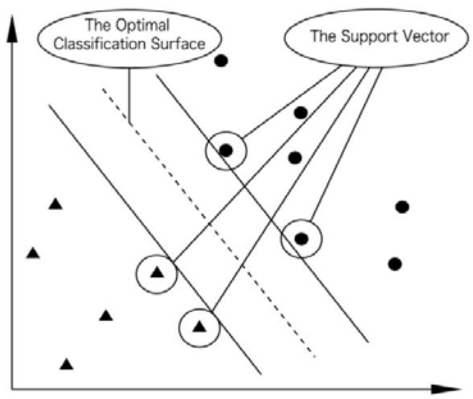

Fig. (1). Linearly separable optimal hyper plane.

For linearly non-separable training samples, with problems such as noise of training samples, considering mistakenly classifying phenomenon, slack variable ξ_i and penalty coefficient C are introduced to solve the problem of hyper plane in formula (1), it can be adjusted as problem below:

min

$$\frac{P\omega P^2}{2}+C\sum_{i=1}^{N}\xi_i$$

s.t.

$$y_i\left(\omega\cdot x_i\right)+b\geqslant 1-\xi_i\ ,\ \xi_i\geqslant 0\ ,$$

i=1…N

Apply multiplicator Lagrange to solve constraint optimal problem:

$$L\left(\omega\text{, b, a, }\xi\right)=\frac{1}{2}\omega^T\omega-\sum_{i=1}^{N}a_i\left[y_i\left(\omega\cdot x_i+b\right)-1+\xi\right] \qquad (3)$$

In the formula (3), ai>0 is Lagrange coefficient. Obtaining the partial derivatives of ω and b, and commanding the results are 0, eventually the question of optimal classification surface turned to the dual problem in finding the maximization of target function.

max

$$Q\left(a\right)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i=1}^{N}a_ia_jy_iy_jx_i^Tx_j$$

s.t.

$$C>a_i>0\ ,\ \sum_{i=1}^{N}a_iy_i=0\ ,\ \text{i=1…N.}$$

The optimal solution of weight vector ω and b is:

$$\omega_0=\sum_{i=1}^{N}a_i^*y_ix_i$$

$$b_0=1-\omega_0x^{(s)}\ ,\ y^{(s)}=1$$

The optimal classification surface is:

$$g\left(x\right)=\omega^Tx+b=\sum_{i}a_iy_ix_i^Tx_j+b$$

x stands for testing samples.

With turning the original problem to dual problem, the calculating complexity is not depended dimension of space. As the default kernel function in LIBSVM toolbox is RBF, the SVM training model is C-SVC model as default model, so the decision function is:

$$f\left(x\right)=\mathrm{sgn}(\omega^Tx+b)=$$
$$\mathrm{sgn}[\sum_{i}a_iy_iK(x_i,x)+b]=$$
$$\mathrm{sgn}[\sum_{i}a_iy_i\exp(-gamma\,\mathrm{P}x_i-x\,\mathrm{P}^2)+b]$$

$\mathrm{P}x_i-x\,\mathrm{P}^2$ is two norm distance, gamma is parameter －g, the default is the reciprocal of number of attributes. Then the default －g is 1/k, the default C is 1.

To obtain a well popularized SVM classification machine, it's the key point to get a optimal penalty parameter C and g of kernel function, Parameter g mostly influenced the distribution complexity of feature space of samples. The penalty coefficient C adjusts the rate between confidence interval and empirical risk in the specific feature space.

2. OPTIMIZATION OF LIBSVM PARAMETERS BASED ON GENETIC ALGORITHM

Genetic algorithm is originated from computer simulation in the biological system. Genetic algorithm simulate the natural selection and duplication, cross and mutation in genetic phenomenon. Starting from either initial population, with selecting, crossing and mutating randomly, a adaptable individual is created, and the region of population turned to a better area, it will converge to the best individual with development from one generation to another, and the optimal solution will be obtained. Genetic algorithm possess a performance of highly parallel ,random and self-adapt searching characteristic, it apparently has a advantage of solving non-linear problem which traditional method could not.

Lithology recognition with logging data is a typical process of non-linear pattern classification. Choosing the parameter plays an important role in the establishing a model of identification of oil shale with the LIBSVM method. In the practical situation, it normally happened that samples are linearly non-separable in the SVM classification, and it could not be separated even after mapping. Therefore choosing the right parameters of LIBSVM is the key to solve the problem. The main processes are shown as below:

1) Binary coding was adopted to encode the chromosome, and 10 bit code represented each parameter, two parameters were shown as 20 bit binary coded string. Top ten code was for penalty coefficient C, the last ten code was for kernel function parameter g. Parameter C was set in the range of(0,C*),C*=max(a$_i$),the searching space for g was [$\min^{(\mathrm{P}x_i - x\,\mathrm{P}^2)\times 10^{-3}}$, $\max^{(\mathrm{P}x_i - x\,\mathrm{P}^2)\times 10^{-3}}$], g$_1$ represented the maximum, g$_2$ represented the minimum.

2) U$_1$ and U$_2$ separately represented decimal integer code of C and g after encoding, corresponding decoding formulas were:

$$C=\left(C^*-0\right)\times\frac{U_1}{2^{10}-1}+0$$

$$g=\left(g_1-g_2\right)\times\frac{U_2}{2^{10}-1}+g_1$$

3) Initial value generated by genetic algorithm was adopted as initial population, the population size was set as 30, big size of population easily caused the increase of the amount of calculation, small size of population could not reflect the diversity of population.

4) Designing the fitness function, assumed that k was the number of cross validation, 50 percent of cross validation was adopted to obtain fitness value of the individual in genetic algorithm.

$$f\left(C,g\right)=\frac{1}{K}\bullet\frac{1}{inaccuracy}$$

Inaccuracy was the rate of error in training samples of LIBSVM, as lower the rate of error of the training samples the higher value of fitness function of chromosome of parameters.

5) In the genetic algorithm, proportional selection operator was adopted in it, one-point crossover was adopted in the crossover operation, the basic mutation operator was adopted in the mutation operation. The termination of the evolution algebra G=100, crossover rate Pc＝0.70, mutation rate Pm=0.03.

6) After repeated calculation of the fitness function value of each population, and operate the genetic operator according to the fitness function value, a new population was generated until it reached the generation 100 or the variation of objective function value never exceed 0.005 and then it would stop the calculation.

3. APPLICATION OF LIBSVM IN RECOGNITION OF OIL SHALE

3.1. Selection of Samples and Normalization

Characteristics of logging curves were researched with geochemical analysis data and log data, it found that logging curves of oil shale indicating the characteristics of high resistivity, high interval transit time, high natural gamma, high neutron porosity and low volume density, Natural gamma GR, logarithm of specific resistivity logR, interval transit time AC and volume density DEN was the input samples, as long as it was oil shale, the output would be 1, the other would be －1. All the output values were shown as column vector. The format of input data was stationary to be double type. If there were data of the other format, it should be changed before it was inputted to calculation. The logging data in y district in Qaidam Basin was chosen as samples. Considering the representativeness and reliability of chosen samples, 500 typical samples were chosen as pattern recognition samples of LIBSVM, and the first 400 were taken as learning samples, the rest of 100 were taken as testing samples to calculate the precision of model.

The method of normalization was usually adopted maximum and minimum in the MATLAB, and which was based on that the maximum and minimum of each feature vector of testing samples was equal to the maximum and minimum of each feature vector of training sample. However it could not satisfy every sample in any condition, and the accuracy after normalization was not smaller than that result of calculation not with normalization. Therefore, normalization was not necessarily, it should be separately treated according to the situations.

Compared the accuracies of various methods of normalization after program operating, they were listed as Table **1** below, which was shown that the accuracy of normalization in the range of [0，1] was the best.

Table 1. Comparison of different method of normalization.

Normalization	Accuracy	Svmtrain Parameters
No Normalization	42.5334%（43/100）	‘－c 2 －g 1’
[－1,1]	96.3801%（96/100）	‘－c 2 －g 1’
[0,1]	98.6425%（99/100）	‘－c 2 －g 1’

3.2. Certification of Parameters C and g Based on Genetic Algorithm

In the svmtrain function of LIBSVM, penalty coefficient C and g could be a distribution value in a certain range. And the C and g which made the highest in accuracy would be the one chosen. The best parameters of SVM would be searched with genetic algorithm. The initial population was generated in a size of 30, and the maximum of population was 100. A optimized SVM which obtained best parameters would be the best calculation model. Then the best parameter c=32.7793,g=4.6352.The accuracy of classification was better as the training numbers increasing, in the meantime the bigger value of the fitness function until it converged stably.

3.3. Application

400 training data were chosen to train LIBSVM, and then predicated the labels with trained model. The comparison results are displayed in the Table **2** it separately represente as

Table 2. Two kinds of prediction results of LIBSVM.

Value of Logging Curves				Prediction Lithology		Actual Lithology
RT	**GR**	**AC**	**DEN**	**LIBSVM**	**GA-LIBSVM**	
15	100	457	2.1482	−1	−1	Non Oil Shale
15	91	431	2.1916	1	1	Oil Shale
17	88	427	2.1985	1	1	Oil Shale
23	88	428	2.1431	1	1	Oil Shale
18	94	320	2.1310	−1	1	Oil Shale
14	99	384	2.1486	1	1	Oil Shale
12	100	373	2.1650	−1	−1	Non Oil Shale
13	99	418	2.1930	−1	−1	Non Oil Shale
16	86	434	2.2310	1	1	Oil Shale
17	82	424	2.1575	1	1	Oil Shale
18	76	433	2.0755	1	1	Oil Shale
17	73	452	1.9907	−1	−1	Non Oil Shale
14	74	434	2.0548	1	−1	Non Oil Shale
16	69	449	2.0840	−1	1	Non Oil Shale
21	64	454	1.9959	−1	−1	Non Oil Shale
20	59	450	2.0036	1	1	Oil Shale
25	59	421	2.0328	1	1	Oil Shale
25	64	395	2.0549	−1	1	Oil Shale
19	71	406	2.0576	1	1	Oil Shale

Table 3. Selection of parameters in LIBSVM and results of recognition.

Numbers	c	g	Accuarcy(%)	Value of Fitness Function
1	129.0788	33.7659	47.7568	18.7549
5	98.2237	24.5407	59.3705	43.9028
10	79.6654	16.8031	72.8829	76.9980
15	58.0956	11.3309	89.7328	127.7867
20	45.2239	7.2246	93.2290	153.2974
25	32.7793	4.6352	97.2806	183.7322
30	32.7793	4.6352	97.2806	183.7322

classification results optimized with the modified LIBSVM and which are not. Final classification accuracy is 97.2806%. According to this result, this method can be applied in the recognition of oil shale. The identification results of y log district in northern Qaidam Basin and the comparison of rock core are shown as Fig. (**2**). Part of classification results of LIBSVM model are shown in Table **3**.

CONCLUSION

With the research of LBSVM in the application of recognition of oil shale in northern Qaidam Basin and the difficulties in choosing a right parameter in SVM, a method to optimize the parameter of LIBSVM, which was based on genetic algorithm, was proposed to do the research. The normalization method was chosen according to the difference of

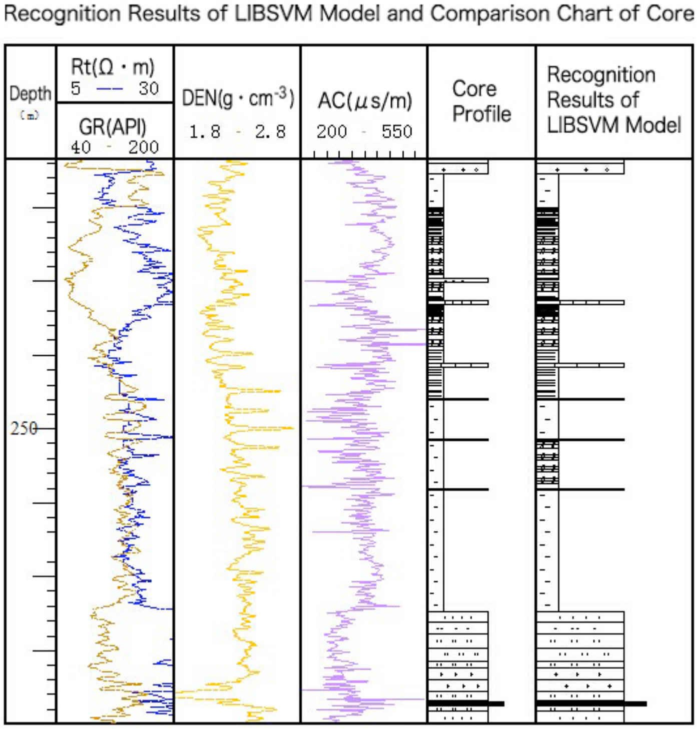

Fig. (2). Recognition results of LIBSVM model and comparison chart of core.

accuracy which obtained by a different normalization method. And optimal penalty coefficient C and parameter of kernel function g were obtained with a modified genetic algorithm, and then this model was used to identify oil shale. The results have shown that the accuracy of recognition of oil shale is reached 97.2806% with a model of modified parameters. And it shows that it a feasible way to deal with the similar problem in the log district in the same geology background. Yet the follow up work is to modify the encoding method and enhance the performance and accuracy of the algorithm.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Geological Exploration Fund Project in Qinghai Province, 2013 (The First Batch) (Land and Resources Department of Qinghai Province [2013] The 64th File).

REFERENCES

[1] J. Zhu, G. Zhao, and B. Liu, "Identification Technology and It's Application of Well-Logging About Oil Shale," *Journal of Jilin University*, vol. 42, no. 2, pp. 289-295, 2012.

[2] J. Zhang, H. Li, F. Yao, "A Geophysical Method for the Identification and Evaluation of Oil Shale," *Acta Petrolei Sinica*, vol. 33, no. 4, pp. 625-631, 2012.

[3] Y. Li, "Study of the Swarm Intelligence Algorithm Based on Optimizing the Parameters of SVM," Tianjin: Tianjin University, 2007.

[4] X. Wang, F. Shi, and Y. Li, "43 Analysis Cases of Neural Network in MATLAB," Beijing: Beijing University of Aeronautics and Astronautics Press, 2013.

[5] X. Yang, Y. Ji, and X. Tian, "Parameters Selection of SVM Based on Genetic Algorithm," *Journal of Liaoning University of Petroleum & Chemical Technology*, ol. 24, no. 1, pp. 54-58, 2004.

[6] Y. Lei, S. Zhang, and X. Li, "Genetic Algorithm Toolbox and Aplication," *Xi'an: Xi'an Electronic Science & Technology University Press*, pp. 3-6, 2014.

[7] L. Liu, and A. Liu, "Fault Diagnosis of Distillation Column Based on Improved Genetic Algorithm Optimization-Based Support Vector Machine," *Journal of East China University of Science and Technology: Natural Science Edition*, vol. 37, no. 2, pp. 228-233, 2011.

[8] Y. Song, J. Zhang, and W. Yuan, "A new Identification Method for Complex Lithology with Support Vector Machine," *Journal of Daqing Petroleum*, vol. 31, no. 5, pp. 18-20, 2007.

[9] X. Zhang, X. Xiao, and L. Yan, "Lithologic Identification Based on Fuzzy Support Vector Machines," *Journal of Oil and Gas Technology*, vol. 31, no. 6, pp. 115-118, 2009.

[10] S. Wang, L. Tao, and H. Wang, "Face Recognition Based on Support Vector Machine and Genetic Algorithm," *Computer Engineering and Application*, vol. 45, no. 12, pp. 164-166, 2009.

[11] Y. Li, "Explaining Videos of SVM," 2010.

Received: April 10, 2015 Revised: May 20, 2015 Accepted: June 15, 2015