Data classification prediction algorithm based on support vector machine and its application in oil and gas reservoir classification

Huaping Yu[1, a], Mei Guo[2, b]

[1] College of Computer Science, Yangtze University, Jingzhou Hubei 434023, China;

[2] Department of Science and Technology, Yangtze University, Jingzhou Hubei 434023, China.

[a]yhpjz@126.com, [b]guomei@yangtzeu.edu.cn

Abstract. For the parameter selection problem of support vector algorithm (SVM) in process of data classification prediction, we put forward the idea of using heuristic algorithms (include genetic algorithm and particle swarm optimization) to optimize the SVM parameters. The principle of genetic algorithm (GA) and particle swarm optimization (PSO) algorithm were analyzed firstly. Secondly, the parameters optimization problem of SVM algorithm with GA and PSO were realized by MATLAB toolbox LIBSVM. Finally, the simulation and experiment based on the data of logging and testing data in an area of Changqing oilfield were discussed on different data normalization ways, different SVM kernel functions and different SVM parameters optimization methods, the results show that the SVM parameters optimization based on the genetic algorithm can significantly improve the accuracy of the algorithm.

Keywords: support vector machine, data classification prediction, parameter optimization, genetic algorithm, particle swarm optimization, oil and gas reservoir classification.

1. Introduction

Support Vector Machine (SVM) is a new type of machine learning algorithms for classification and regression problems based on the statistical learning theory [1], which is first proposed by Vapnik. [2]. In recent years, the data classification prediction algorithm based on SVM is widely used in all kinds of data classification prediction analysis [3,4,5]. Selection of SVM learning parameters has a great influence on its performance, which only relies on the experience of model builder. Therefore, this paper adopts the genetic algorithm (GA) and particle swarm optimization (PSO) to optimize the SVM parameters, and two new GA-SVM and PSO-SVM models were established. Moreover, oil and gas reservoir is divided into various types, which includes oil layer, water layer, gas layer and dry layer etc. And the accurate identification of reservoir types is the basic demand of the oil and gas development [6].We used the GA-SVM and PSO-SVM to carry out reservoir classification prediction based on the logging data (such as resistivity, porosity, permeability, shale content, oil saturation, acoustic time, etc.) and testing data, which achieved good results.

2. The Principle of SVM

The main principle of SVM is to create a classification hyper-plane as a decision surface, which makes the edge of the isolation between the positive cases and the negative cases been maximized; The prominent characteristics of SVM is the statistics of structural risk minimization (SRM) principle [7], which achieves the minimization of empirical risks and confidence risks and make it has strong generalization ability. A specific form of typical binary classification SVM is as follows:

1) Setting the known training set:

$$T = \{(x_1, y_1), ..., (x_l, y_l)\} \in (X \times Y)^l \quad (1)$$

Among equation (1), $x_i \in X = R^n, y_i \in Y = \{1, -1\}(i = 1, 2, ..., l)$, x_i is the feature vector.

2) Selecting the proper kernel function k(x, x') and parameter C, and then solving the optimization problem:

$$\min_{a} \frac{1}{2}\sum_{i=1}^{j}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \partial_j \text{ , s.t. } \sum_{i=1}^{l} y_i \alpha_i = 0, 0 \le \alpha_i \le C, i = 1,...,l \quad (2)$$

The optimum solution is $\alpha^* = (\alpha_1^*, ..., \alpha_2^*)^T$.

3) Selecting a positive component ($0 < \alpha_j^* < C$) of α^* is, and then calculating the threshold:

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(x_i - x_j) \quad (3)$$

4) Structuring the decision function:

$$f(x) = \text{sgn}(\sum_{i=1}^{l} \alpha_i^* y_i K(x, x_i) + b^*) \quad (4)$$

SVM algorithm was originally designed for binary classification problems. When dealing with multiple classification problems, a suitable multiple classifiers need to be constructed. Currently, the methods of constructing the SVM methods mainly have two kinds: one kind is the direct method, which directly improves in the objective function, and the multiple classification parameters is merged into an optimization problem. This method looks like simple, but its computational complexity is high and it is difficult to be implemented, which is only suitable for small-scale problems. Another kind is the indirect method, which combines multiple binary classification classifiers to achieve the structure of multiple classifiers. The typical architecture of SVM is shown in figure 1.

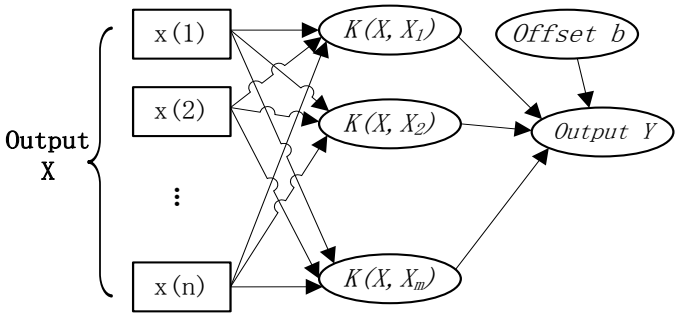


Fig. 1 The architecture of SVM

2.1. Four main kernel functions.

In figure 1, K denotes kernel function. SVM is fully described by the training set and kernel function. Therefore, how to construct and the choice the kernel function is an important problem. The categories of kernel function are as follows:

Linear kernel function: $K(x, x_i) = x^T x_i$

Polynomial kernel function: $K(x, x_i) = (\gamma x^T x_i + r)^p, \gamma > 0$

Radial basis kernel function: $K(x, x_i) = \exp(-\gamma \| x - x_i \|^2), \gamma > 0$

Sigmoid kernel function: $K(x, x_i) = \tanh(\gamma x^T x_i + r)$

2.2 The flow of data classification prediction and data preprocessing techniques.

The general process of data classification prediction based on SVM is shown in figure 2, which includes four steps: determining the training data set and testing data set, data preprocessing, training the SVM model, data classification prediction.

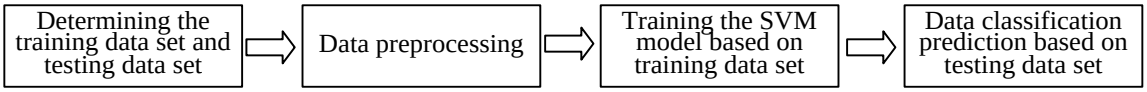


Fig. 2 The flow of data classification prediction

The normalization preprocessing method for the training data set and testing data set is as follows,

$$f: x \to y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{5}$$

In equation (5), $x, y \in R^n$; $x_{\min} = \min(x)$; $x_{\max} = \max(x)$, $y_i \in [0,1], i = 1,2,l,...,n$, That is named as [0,1]. In addition to normalized way [0, 1], there is normalization method [-1, 1], which is maped as below,

$$f: x \to y = 2 \times \frac{x - x_{\min}}{x_{\max} - x_{\min}} + (-1) \tag{6}$$

In MATLAB, mapminmax function can realize the normalization. The common functions interface [8] is as follows,

[y,ps] = mapminmax(x); [y,ps] = mapminmax(x,y_{min},y_{max}); [x,ps] = mapminmax('reverse',y,ps);

Among them, x is the raw data, y is the data after normalization, ps is a structure, which records the mapping of normalization. The mapping of mapminmax function is as follows,

$$y = (y_{\max} - y_{\min}) \times (x - x_{\min}) / (x_{\max} - x_{\min}) + y_{\min} \tag{7}$$

Among them, x_{min} and x_{max} are the minimum and maximum raw data respectively, y_{min} and y_{max} are the range of parameters of mapping, which can be adjusted, and the default values are -1 and 1, that is the [1, 1] mapping normalization function. If y_{min} and y_{max} are set with 0, 1, which is be known as [0, 1] mapping normalization function, the code is as follows,

[y,ps] = mapminmax(x); ps.y_{min} = 0; ps.y_{max} = 1; [y_{new},ps] = mapminmax(x,ps);

The non-normalization can be implemented with the following code:

[y,ps] = mapminmax(x); [x,ps] = mapminmax('reverse',y,ps);

3. SVM Parameter Optimization

Training SVM model need to adjust the relevant parameters (mainly include the penalty parameter c and kernel function parameter g) to get ideal forecast classification accuracy. In figure 2, the svmtrain function in MATLAB generally is used to train the SVM model, but the penalty parameter c and kernel function parameter g of svmtrain are determined by the test experience, thus its parameters are not optimal [8,9], which seriously the accuracy of predicted results. Then, how to select the parameters c and g? Using cross validation (CV) method and the heuristic algorithm (such as genetic algorithm and particle swarm optimization) to choose the best parameters c and g is the effective way to improve the accuracy of the data classification prediction based on SVM.

3.1 SVM parameters optimization based on cross validation method.

Cross validation (CV) is a statistical analysis method used to verify the performance of classifiers [7]. And k-CV method is a typical cross validation method. Namely, the raw data were divided into k groups divide (dividing equally), and each group do one time cross validation and establish their model respectively, and then get k SVM models based on the k data subsets. We can use the average classification accuracy of the final validation set of the k model as the performance index of k-CV classifier. K is generally greater than 2, and k is usually begin from 3 in the actual operation. Only in a small amount of raw data, k will try to get 2. The k-CV can effectively avoid the happening of learning and owe learning state. Finally the result is more convincing.

The classification accuracy with k-CV is trained and verified under condition of each sub-group parameters c and g. Finally, the parameters c and g in highest classification accuracy of k group is known as best parameters c and g. When there may be multiple sets of c and g corresponds to the highest classification accuracy, the group c and g in the condition of the smallest parameter c are known as the best parameters. If there multi-groups parameters c and g to corresponding minimum c, the first group of c and g is selected as the best argument. The reasons for doing so is that too high parameter c will lead to a learning state, that is the training set classification accuracy is high and the test set classification accuracy rate is very low (classifier's generalization ability reducing), therefore

smaller penalty parameter c is a better choice when the classification accuracy reach the highest precision. The pseudo code of parameters selection algorithm based on cross validation method is as follows:

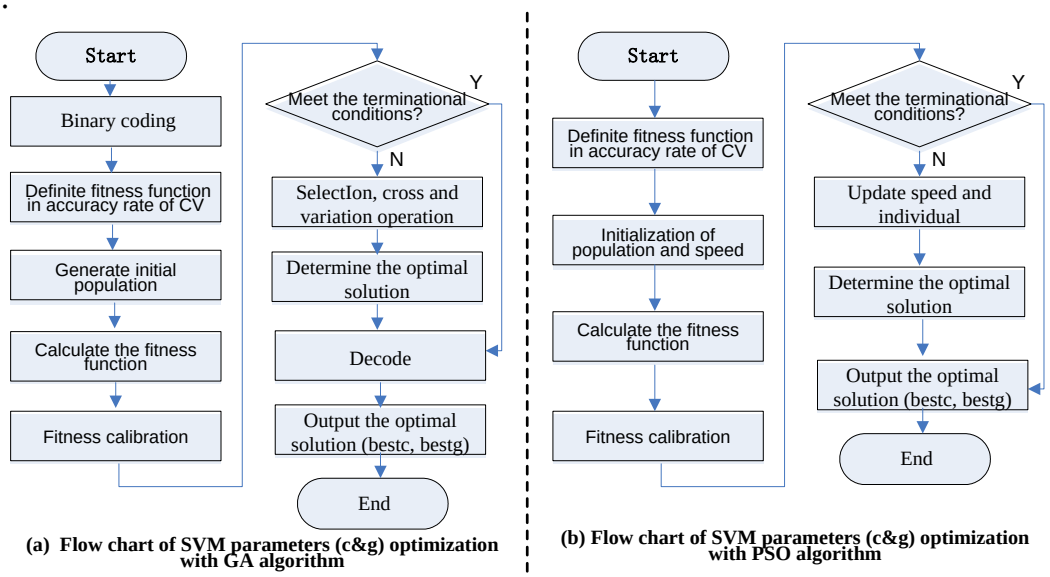
```
Start
% The corresponding data initialization.
bestAccuracy = 0; bestc = 0; bestg = 0;
% Dividing the c and g into grid and search optimal value.
for c = 2^(cmin):2^(cmax)
    for g = 2^(gmin):2^(gmax)
%The training data set is roughly averagely divided into k groups based on k-CV method, which are denoted as r train(1), train(2),…, train(k). The corresponding labels are denoted as train_label(1), train_label(2),…, train_label(k).
        for run = 1:K
        Using train(run) as a training set, the other data set is recorded as the testing set. Using acc(run) to record the verification accuracy.
        end
        cv = (acc(1)+acc(2)+...+acc(K))/k;
        if( cv>bestAccuracy)
            bestAccuracy = cv;      bestc = c;      bestg = g;
        end
    end
end
```

3.2 SVM parameters optimization based on genetic algorithm.

Genetic algorithm (GA) originates in the computer simulation of biological system [9]. 1967, Bageley proposed the word "genetic algorithm" in his doctoral thesis for the first time, he developed the genetic operator such as the copy, mutation, dominant, inversion and diploid coding method. etc. and GA is a kind of practical, efficient and robust optimization techniques, its development is very rapid and GA has aroused great attention of scholars both at home and abroad. The flow chart of SVM parameters (c and g) optimization with GA algorithm is shown in figure 3(a).

3.3 SVM parameters optimization based on particle swarm optimization.

The search space problem is viewed as birds flying space in particle swarm optimization (PSO) algorithm. And every bird feeding in the air is abstract to a particle of search optimal solution in the solution space. The process of birds are looking for food is the optimal solution of optimization problem [10]. The algorithm is put forward by Kenned and Eberhart in 1995 at the earliest. The overall process of SVM parameters (c and g) optimization with PSO algorithm is shown in figure 3(b) [10,11].



(a) Flow chart of SVM parameters (c&g) optimization with GA algorithm

(b) Flow chart of SVM parameters (c&g) optimization with PSO algorithm

Fig. 3 The flow chart of SVM parameters (c&g) optimization with GA algorithm and PSO algorithm

4. Algorithm Simulation and Applications

This section adopts logging and oil test data of 100 small layers of a well block in Changqing oilfield as the data set (the structure of the data set is as shown in table 1). The simulation analysis based on the LIBSVM tool of MATLAB is presented, the result show that different kernel function, data preprocessing methods and SVM parameter optimization methods have a significant impact on data classification prediction accuracy.

Table 1 Logging and oil test data of 100 small layer of a well block in Changqing oilfield (excerpts)

No.	Reservoir interval(m)	Reservoir logging well and testing parameters						
		Resistivity Ω·m	Porosity %	Permeability mD	Oil saturation %	AC us/m	Shale content %	Test results
1	2229.8-2232.3	30.84	11.41	0.24	47.26	224.98	14.01	oil layer
2	2233.7-2236.1	29.31	10.49	0.18	42.95	219.17	16.53	oil layer
3	2237.6-2241.2	27.6	11.92	0.27	45.86	228.25	14.34	oil layer
…	…	…	…	…	…	…	…	…
94	2163.3-2172.2	37.35	10.94	0.22	46.35	222.07	16.79	poor oil layer
95	2172.6-2174.1	38.27	10.03	0.15	44.57	216.89	18.01	dry layer
96	2183.9-2186.3	53.42	11.27	0.23	56.38	224.80	16.18	water layer
97	2187.5-2190.5	55.78	10.32	0.17	54.75	217.84	15.37	oil layer
98	2190.5-2194.6	48.26	10.05	0.16	50.71	216.37	17.80	poor oil layer
99	2200.8-2202.8	47.55	9.51	0.13	35.65	212.22	13.85	dry layer
100	2204.1-2205.5	45.39	10.41	0.18	36.91	219.86	17.71	dry layer

4.1 The influence of different kernel functions on prediction accuracy

Kernel functions of SVM include linear kernel function, polynomial kernel function, the radial basis kernel function and sigmoid kernel function, which is described in section 2.1. The influence of different kernel functions on data classification prediction accuracy based on the unified the normalization [0, 1] is as shown in figure 4. In figure 4, the radial basis function has the highest classification accuracy.

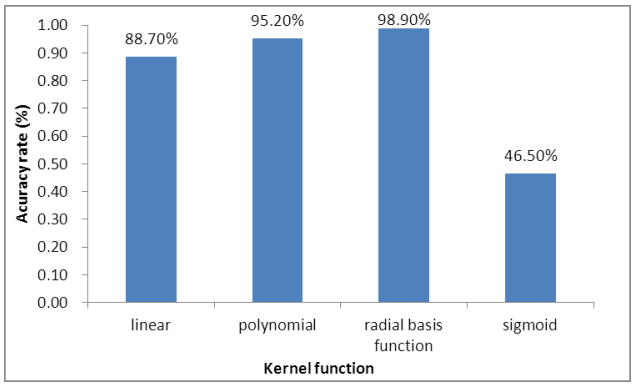


Fig. 4 Data classification prediction accuracy comparison based on different kernel function of SVM ([0, 1])

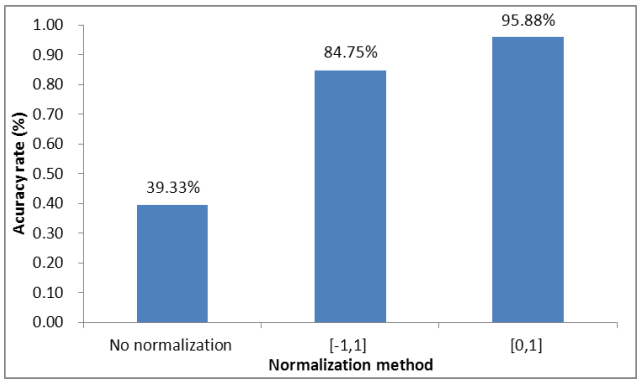


Fig. 5 Data classification prediction accuracy comparison based on different data normalization methods of SVM

4.2 The influence of different normalization methods on prediction accuracy.

Data normalization methods of SVM include [0,1] method and [-1,1] method, which is described in section 2.2. The influence of different data normalization methods on data classification prediction accuracy based on radial basis kernel function is as shown in figure 5. In figure 5, different ways of normalization have certain effect on the final accuracy. Specifically, the logging well and test data an area of Changqing oilfield need to be normalized, especially the [0, 1] normalization method can improve the accuracy of the final classification.

4.3 The influence of different SVM parameters optimization method on prediction accuracy.

The svmtrain function in MATLAB, GA algorithm and PSO algorithm were used respectively to optimize the SVM parameters. The results show that the SVM algorithm based on GA has better accuracy for data classification prediction with the logging and testing data of 100 small layers in Changqing oilfield, which are shown in figure 6.

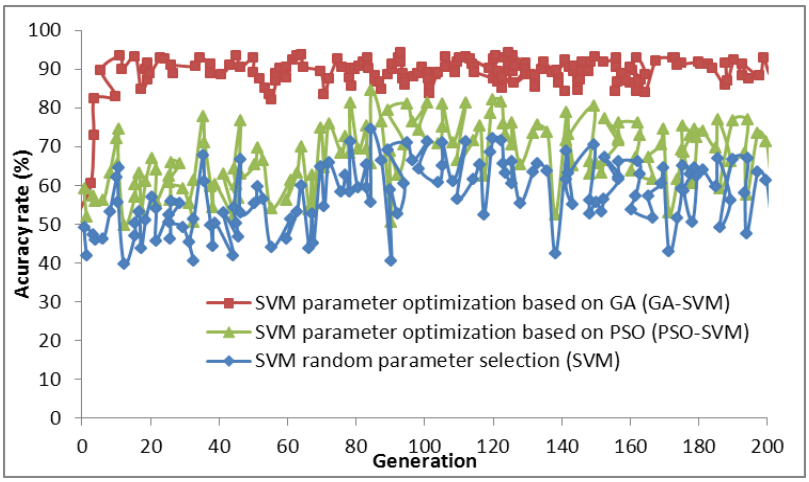


Fig. 6 Data classification prediction accuracy comparison based on different SVM parameters optimization methods

As can be seen from the above experiment and comparison results, the set up GA-SVM algorithm has higher prediction precision and adaptability for a well area of Changqing oilfield. GA-SVM model has a better calculation results compared with PSO-SVM algorithm and simple SVM algorithm, which has proved that GA-SVM algorithm has feasibility and effectiveness in reservoir classification prediction.

5. Conclusion

This paper selects 100 reservoir samples as the training and test data set. The training data set is used to train SVM model, and the test data set is applied to reservoir classification prediction. And the SVM parameter optimization methods, SVM kernel functions and data normalization methods are detailed Analyzed, the result shows that the SVM classification of prediction model can effectively realize the identification of reservoir category, the GA-SVM has more performance compared with CV and PSO algorithms.

Acknowledgements

All of the authors are grateful to the anonymous referees for their valuable comments and suggestions. The research is supported by the PetroChina Innovation Foundation (Grant No.2013D-5006-0605).

References

[1] Cortes C and Vapnik V: Support-Vector network, Machine Learning, Vol. 20 (1995), p. 273-297.

[2] Boser B, Guyon I. and Vapnik V: A training algorithm for optimal margin classifiers, *ACM press: In Proceedings of the Fifth Annual Workshop on Cornputetional Learning Theory*, 1992.

[3] HSU C W and LIN C J．: A simple decomposition method for support vector machines, Machine Learning, Vol.46 (2002), No. 1, p. 291-314.

[4] Tran Quang-anh, ZHANG Qian-li and LI Xing: SVM classification-based intrusion detection system, Journal of china institute of communications, Vol. 23 (2002), No.5, p. 51-56.

[5] FAN Xin-wei, DUS hu-xin and WU tie-jun :Rough support vector machine and its application to wastewater treatment processes, Control and Decision, Vol.19 (2004), No.5, p.573-576.

[6] Z.Y. Liu, Y. Wang and R.R. Cao: Study of the reservoir classification by logging evaluation, World well logging technology,Vol. 23 (2008) No.4, p 19-22.

[7] V.Vapnik: *The Nature of Statistical Learning Theory* (Tsinghua University Press, China Beijing 2000.9).

[8] Lei yinjie, Zhang shanwen and Li xuwu: *MATLAB genetic algorithm toolbox and its application* (Xian University of electronic science and technology, China Xian 2005).

[9] Yang Jie, Zheng Ning, Liu Dong and, Luo Shigui: Optimization of Features with Weight and Model Parameters of SVM Based on Genetic Algorithm, Computer Simulation, Vol.25(2008), No. 9, p.113-118.

[10] Zeng jiancao, Jie jin and Cui zhihua: *Particle swarm optimization* (Science Press, China Beijing 2004).

[11] Zhang Qing and Liu Bingjie: Fast Optimization Method for Parameter of SVM Based on PSO and Divided Training, Science Technology and Engineering, Vol.8(2008), No. 16, p.4513-4616.