Pet.Sci.(2010)7:379-384
DOI 10.1007/s12182-010-0081-1

Using a support vector machine method to predict the development indices of very high water cut oilfields

Zhong Yihua[1]*, Zhao Lei[1], Liu Zhibin[1], Xu Yao[2] and Li Rong[1]

[1] School of Sciences, Southwest Petroleum University, Chengdu, Sichuan 610500, China
[2] Sichuan Forestry Cadre School, Chengdu, Sichuan 610066, China

Abstract: Because the oilfields in eastern China are in the very high water cut development stage, accurate forecast of oilfield development indices is important for exploiting the oilfields efficiently. Regarding the problems of the small number of samples collected for oilfield development indices, a new support vector regression prediction method for development indices is proposed in this paper. This method uses the principle of functional simulation to determine the input-output of a support vector machine prediction system based on historical oilfield development data. It chooses the kernel function of the support vector machine by analyzing time series characteristics of the development index; trains and tests the support vector machine network with historical data to construct the support vector regression prediction model of oilfield development indices; and predicts the development index. The case study shows that the proposed method is feasible, and predicted development indices agree well with the development performance of very high water cut oilfields.

Key words: Oilfield development indices, oilfield performance, support vector regression, high water cut, time series

1 Introduction

In most oil fields in eastern China, the water cut of produced fluids exceeds 90%. Such oilfields are called very high water cut oilfields. It becomes more and more difficult to stabilize the oil production of these oilfields (Sun, 2006) since the distributions of oil and water have greatly changed in the very high water cut reservoir. During this production phase, the oil production decreases sharply and the remaining oil is believed to be widely dispersed over large areas of the pay. It is very difficult to recover the remaining oil in the reservoir. Therefore, such measures as selective water injection, selective hydraulic fracturing, and selective water shutoff, are performed in high water cut reservoirs to control the increasing rates of water injection and liquid production, and to enhance oil recovery (Zhong, 2009).

Since the relationship between oilfield performance (development indices) and its influencing factors is generally nonlinear during very high water cut production, it is very difficult to establish an analytical model for predicting development indices. Many oil reservoirs in eastern China are producing at very high water cut, but the amount of data available about oilfield performance and its influencing factors is very small. Therefore, the regression analysis method (Wang and Chen, 2004), grey prediction method (Yao et al, 2007), differential simulation, and neural network simulation (Chen and Lang, 2003; Liu et al, 2008), which are based on a large number of cases, cannot be used to predict the development indices due to the small sample size of cases. However, the small sample size and nonlinear prediction problem can be solved by a support vector machine (SVM) method (Zhang, 2004). Zhong (2009) proposed that an SVM is a very effective method for predicting development indices of very high water cut oilfields.

In this study, we use an SVM method to predict the development indices of very high water cut oilfields, and a new method to select its kernel function is proposed.

2 Theory of support vector regression and its improvement

2.1 SVR principles

The support vector machine (SVM) method was originally proposed for pattern recognition. When Vapnik introduced the □–insensitive loss function to SVM, SVM was extended to support vector regression (SVR) which has excellent performance in the nonlinear regression estimation problems of small sample size. Its basic idea is that the input variables

*Corresponding author. email: zhongyh_65@126.com
Received June 23, 2009

(training sample vector) are mapped from the primal space to higher dimensional feature space based on the information of given training samples and through nonlinear mapping $\varphi(x)$ to construct a linear decision-making (regression) function to realize the linear regression in this feature space (Zhang, 2004).

Support vector regression (SVR) is a powerful technique for solving the regression problem, which is used to estimate a regression function $f(x)=w^{\mathrm{T}}\cdot\varphi(x)+b$ to minimize the expected risk $R[f]=\int c(x,y,f)\,\mathrm{d}\,p(x,y)$, and to predict accurately the output y corresponding to a new sample point x according to a small number of given training sample points $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(x,y)^l$, $x_i\in x=R^n$, $y_i\in y=R$, i=1,$\cdots$, l, and a given loss function $c(x,y,\cdots,f(x))$. x_i denotes the input vector; y_i denotes the output (target) value; l denotes the total number of data samples (Zhang, 2004; Cheng et al, 2007; Goh and Goh, 2007).

2.2 A new method for selecting the SVR kernel function

The generalization performance of SVR depends largely on the selection of the kernel function. There is so far no general theoretical basis for the selection of kernel function. The optimum kernel function is generally found using the method of exhaustion for the most common kernel functions in the literature, which is somewhat frustrating and time-consuming and does not make full use of the input information about the regression prediction problem (Ito and Nakano, 2003; Feng and Yang, 2007). Hence, to solve the nonlinear regression problem of small sample time series, a new method for selecting the SVR kernel function is proposed in this paper. It is based on the statistical analysis of a set of historical data of prediction indices. This method first analyzes some time series characteristics of the prediction indices, such as tendency, seasonal nature, periodicity, and randomness; then selects the kernel function according to the time series characteristics of prediction indices. Based on the basic idea of SVR and the characteristics of the kernel function, the nonlinear mapping $\varphi(x)$ is replaced by the kernel function $k(x_i,x_j)=\varphi(x_i)\varphi(x_j)$ to compute the prediction indices (Zhang, 2004). Therefore, it is wise to select the type of kernel function according to the type characteristics of the time-series graph of prediction indices (Box et al, 2005; Wang and Hu, 2007) and the kernel function graph to ensure that the calculated values of prediction indices follow a similar trend to the historical data. Suppose that $k:X\times X\to R$ is a kernel function, Φ is the feature mapping of X, k makes a pseudo-distance $\rho_k(x,x')=\|\Phi(x)-\Phi(x')\|$ in input space X (Wang, 2006). This pseudo-distance can be interpreted as a measure of similarity between x and x'. Selecting the kernel function is equivalent to defining the similarity of elements in input space, so it is necessary to consider the characteristics of the problems while selecting a kernel function (Chapelle et al, 2002b); Momma and Bennett, 2002; Cherkassky and Ma, 2003). If the time series graph of prediction indices shows trend variation, then a linear kernel function or polynomial kernel function may be selected as the SVR kernel function. If the time series graph of prediction indices shows periodic movement or variation, seasonal movement or variation, then a sigmoid kernel function may be selected as the SVR kernel function (Lin and Lin, 2003). If the time series graph of prediction indices shows irregular or random movement, then a radial basis function (RBF) kernel may be selected as the SVR kernel function (Steinwart, 2002).

3 The SVR model and method for predicting development indices of very high water cut oilfields

3.1 SVR model

When the oilfield (or reservoir) is modeled or treated as a system, its complicated internal mechanisms are neglected, and oilfield development is viewed as an input-output process, development indices and their influencing factors form a complicated input-output system (Li and Liu, 2001). From the point of view of system theory, various complexity factors of the whole oilfield or reservoir closely interact, so it is difficult to establish a prediction model $y(t)=f(x(t))$ between development indices $y(t)$ and their influencing factors $x(t)\in R^n$, which can include the field performance. Because of the complexity of reservoir geology and the small amount of oilfield production data collected, the development system of a water-drive oilfield with a very high water cut is first treated as a whole. Then the input-output system of the SVR network model which can indicate the oilfield performance is established according to functional simulation of the oil-gas system (Li and Liu, 2001) and the change information of past state of oil-gas system, i.e. historical data on oilfield production. The SVR network model is trained using historical data on oilfield production. Then functional isomorphism is realized with the trained SVR network model. Finally the SVR network model is extrapolated to predict the future change of the system to provide guidance for controlling decision-making.

3.2 SVR algorithm

According to the analysis mentioned above, the SVR algorithm for development indices of very high water cut oilfields is as follows.

Step 1 Determine the input and output of SVR prediction system

Establish the development indices and the important influencing factors using the correlation analysis method. The development indices and influencing factors are treated as the input and output of SVR prediction system.

Step 2 Construct the SVR prediction model

1) Determine and normalize the training sample set $T=\{(x_i,y_i)|i=1,2,\cdots,l\}$ according to Step 1, where x_i denotes the ith sample of the factor set that reflects the oilfield development index; y_i denotes the ith sample of the oilfield development index.

2) Select design parameters of SVR, such as C, ε, and kernel function $K(x,x')$.

3) Construct the optimization problem

$$\begin{aligned}&\min_{\alpha^{(*)}\in R^{2l}} \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)+\\&\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i),\\&s.t.\quad \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0,\\&\qquad 0\le\alpha_i,\quad \alpha_i^*\le\frac{C}{l},\quad i=1,2,\cdots,l\end{aligned}\tag{1}$$

where α_i, α_i* are Lagrange multipliers, which satisfy the equalities $\alpha_i\alpha_i$*=0; C is a penalty factor, and a larger value of C means a larger penalty value of fitness bias; and ε is a positive constant and the largest allowable error of regression, whose value is selected in advance.

4) Solve Eq. (1) to obtain optimal Lagrange multipliers $\overline{\alpha}=(\overline{\alpha}_1,\overline{\alpha}_1^*,\cdots,\cdots\overline{\alpha}_l,\overline{\alpha}_l^*)^{\mathrm{T}}$.

5) Construct a regression function, which makes the expected risk $R[f]$ minimum

$$f(x)=\sum_{i=1}^{l}(\overline{\alpha}_i^*-\overline{\alpha}_i)K(x_i,x)+\overline{b}\tag{2}$$

where the samples x_i corresponding to non-zero $\overline{\alpha}_i^*-\overline{\alpha}_i$ are support vectors; $\overline{w}=\sum_{i=1}^{l}(\overline{\alpha}_i^*-\overline{\alpha}_i)K(x_i,x)$; $\overline{b}$ is calculated as follows: select $\overline{\alpha}_j$ or $\overline{\alpha}_k^*$ from $(0,\frac{C}{l})$, if $\overline{\alpha}_j$ is selected, then $\overline{b}=y_j-\sum_{i=1}^{l}(\overline{\alpha}_i^*-\overline{\alpha}_i)K(x_i,x_j)+\varepsilon$; if $\overline{\alpha}_k^*$ is selected, $\overline{b}=y_k-\sum_{i=1}^{l}(\overline{\alpha}_i^*-\overline{\alpha}_i)K(x_i,x_k)-\varepsilon$.

6) Test the regression/decision function using the test data set. If the desired prediction accuracy is achieved, the prediction model of the development index is obtained; otherwise modify and adjust the design parameters and the kernel function (Chapelle et al, 2002a); Chung et al, 2003; Zhu et al, 2004; Lin et al, 2008).

Step 3 Predict the development index by the prediction model generated in Step 2

Input or predict the influencing factors of the development index using the time series method (Box et al, 2005), and predict the development index by the generated development index prediction model.

4 Case study

4.1 Calculation

The SVR method proposed in Section 3 was used to predict oil production and liquid production of integrated oilfield A2, the Shengli Oil Field, China in its very high water cut stage. Other development indices can be directly calculated from these two indices.

According to Step 1 in Section 3.2, the factors influencing the monthly oil production include the remaining reserves (Q_{ro}) in this oilfield, the total number of oil production wells (N_o), monthly injection-production ratio (R_{IP}), water cut (f_w), the number of active water injection wells (N_{wa}), the number of new production wells (N_{on}), and the number of effective stimulation treatments for old wells (N_t). There are also seven factors influencing the monthly liquid production. A total of 12 data sets of oil production, liquid production, and their influencing factors selected from December 1993 to November 1994 are listed in Table 1.

Table 1 Historical data on oil production, liquid production, and their influencing factors of integrated oilfield A2

Time	Q_{ro} 10^8 tonnes	N_o Wells	R_{IP}	f_w %	N_{wa} Wells	N_{on} Wells	N_t Times	Oil production 10^6 tonnes	Liquid production 10^6 tonnes
1993-12	21.74	1329	1.08	90.25	666	124	553	0.332	3.408
1994-01	22.01	1342	1.10	90.36	660	13	26	0.314	3.261
1994-02	21.99	1353	1.10	90.43	663	24	53	0.284	2.966
1994-03	21.95	1359	1.09	90.41	670	33	75	0.315	3.288
1994-04	21.93	1383	1.09	90.59	665	58	114	0.296	3.143
1994-05	21.89	1396	1.09	90.65	662	72	155	0.305	3.268
1994-06	21.87	1399	1.09	90.62	674	80	193	0.293	3.121
1994-07	21.84	1408	1.09	90.91	682	90	214	0.297	3.272
1994-08	21.81	1408	1.11	91.17	678	93	239	0.297	3.358
1994-09	21.78	1407	1.13	91.69	680	92	261	0.287	3.453
1994-10	21.75	1286	1.15	92.02	685	96	277	0.288	3.605
1994-11	21.72	1292	1.16	92.03	692	105	297	0.278	3.494

The historical data on oil production, liquid production, and their influencing factors were respectively normalized by $z_{\text{normal}} = \dfrac{z - z_{\min}}{z_{\max} - z_{\min}}$, where $z_{\min}$, $z_{\max}$ are the minimum and maximum in the data to be normalized respectively. Three data sets selected randomly such as the data sets of July, September, and November in 1994 were taken as a testing set; and the other nine data sets were taken as a training set. Because the oil production and liquid production varied randomly with time, the RBF kernel function was selected using time series analysis. Parameters $\varepsilon = 0.006$, $C = 35.3887$ were selected by the empirical formula (Feng and Yang, 2007). The Kernel function parameter $\sigma^2=0.05$ was determined by a 3-fold cross-validation of the training set (Ito and Nakano, 2003). The nine training samples were used to train the SVR model. The mean absolute percentage error (MAPE) and root mean square error (RMSE) are used to evaluate prediction accuracy, which are as follows:

$$MAPE = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{y_k - \hat{y}_k}{y_k}\right| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2} \tag{4}$$

where y_k and $\hat{y}_k$ represent the actual and predicted values, respectively; N is the number of prediction points.

The MAPE and RMSE of SVR of nine training samples for oil production are 0.352% and 0.140, respectively; those for liquid production are 0.526% and 0.206, respectively. The number of support vectors is six. Three test samples were used to evaluate the trained SVR model. The MAPE and RMSE of SVR of three test samples for oil production are 0.767% and 0.236, respectively; those for liquid production are 0.289% and 0.229, respectively. Table 2 shows that the predicted oil production and liquid production of the integrated oilfield A2 in December 1994 are separately 0.278×10^6 tonnes and 3.659×10^6 tonnes.

Table 2 Monthly oil production and liquid production (10^6 tonnes) of the integrated oilfield A2 and the MAPE (%) and RMSE values predicted by different methods

Time		SVR		FSBOTVS		HDCA	
		Oil production	Liquid production	Oil production	Liquid production	Oil production	Liquid production
1994-12		0.278	3.659	0.278	3.647	0.279	3.483
1995-01		0.276	3.607	0.276	3.587	0.276	3.469
1995-02		0.274	3.510	0.278	3.499	0.272	3.395
1995-03		0.274	3.411	0.278	3.405	0.269	3.338
1995-04		0.272	3.351	0.277	3.339	0.266	3.289
Training	MAPE	0.352	0.526	0.382	0.752	6.896	4.018
	RMSE	0.140	0.206	0.154	0.281	2.340	1.487
Testing	MAPE	0.767	0.289	0.835	0.607	8.961	4.126
	RMSE	0.236	0.229	0.258	0.298	2.285	1.527

4.2 Case analysis

In order to validate the SVR method, hyperbolic decline curve analysis (HDCA) and functional simulation based on time-varying system (FSBOTVS) (Liu et al, 2008) were used to predict the oil production and liquid production from December 1994 to April 1995 of the integrated oilfield A2 besides the method presented in this paper. Predicted results are listed in Table 2. The MAPE and RMSE values in Table 2 show the generalization (forecast) ability of methods used. The lower the MAPE and RMSE values are, the better the predicted result is. The oil production and liquid production of the fault block oilfield B4 (32 samples, small sample size) and the integrated oilfield A3 (112 samples, relatively large sample size), Shengli Oil Field, China were also predicted by different methods. Results are listed in Tables 3 and 4. The predicted results indicate that the generalization ability of SVR and FSBOTVS are good. The predicted results agree well with the actual values, and reflect the trend of output. While the prediction accuracy of HDCA is relatively low and the results predicted with this method do not reflect the output fluctuation. This further indicates that HDCA is not suitable for prediction of development indices of very high water cut oilfields. SVR is better than other prediction methods for the small sample size problems; its prediction accuracy is almost the same as that of FSBOTVS for relatively large sample size problems. However, the prediction accuracy is lower than that of FSBOTVS when predicting multiple values of problems.

Table 3 Monthly oil production and liquid production (10^4 tonnes) of fault block oilfield B4 and the MAPE (%) and RMSE values predicted by different methods

Time		SVR		FSBOTVS		HDCA	
		Oil production	Liquid production	Oil production	Liquid production	Oil production	Liquid production
2003-10		0.47	4.60	0.47	4.67	0.58	5.16
2003-11		0.39	3.80	0.39	3.83	0.58	5.16
2003-12		0.36	4.20	0.36	4.26	0.57	5.16
2004-01		0.38	4.63	0.38	4.68	0.57	5.16
2004-02		0.34	4.12	0.34	4.24	0.57	5.15
Training	MAPE	3.24	2.01	3.54	3.45	22.21	20.69
	RMSE	0.017	0.114	0.018	0.144	0.119	1.478
Testing	MAPE	4.29	1.67	4.64	2.65	29.91	20.59
	RMSE	0.200	0.097	0.212	0.159	1.391	1.237

Table 4 Monthly oil production and liquid production (10^6 tonnes) of integrated oilfield A3 and the MAPE (%) and RMSE values predicted by different methods

Time		SVR		FSBOTVS		HDCA	
		Oil production	Liquid production	Oil production	Liquid production	Oil production	Liquid production
1999-12		0.252	4.293	0.252	4.293	0.253	5.069
2000-01		0.256	4.386	0.256	4.354	0.255	5.078
2000-02		0.260	4.339	0.260	4.357	0.253	5.087
2000-03		0.259	4.302	0.259	4.308	0.252	5.097
2000-04		0.257	4.383	0.257	4.329	0.251	5.106
Training	MAPE	0.59	0.38	0.53	0.36	3.210	4.05
	RMSE	0.220	2.157	0.198	2.054	1.325	23.31
Testing	MAPE	1.15	0.45	1.18	0.44	2.88	9.46
	RMSE	0.312	2.338	0.319	2.275	0.990	45.39

5 Conclusions

Prediction of oilfield performance is important for the design and development of oil fields. Based on analysis of conventional prediction methods of development indices and factors influencing the oilfield performance, an SVR method is established to predict the development indices of very high water cut oilfields. The new SVR method is based on time series analysis to select the kernel function. This method takes into account the correlation between development indices and their influencing factors as well as the prediction accuracy and reliability of the model; it can overcome some shortcomings existing in commonly-used prediction methods and can simultaneously predict the development indices under complex production conditions. The case study shows that the results predicted by the SVR model can reflect the dynamic characteristics of different oil reservoirs at the development stage, and can provide a theoretical basis and important technical support for making and establishing scientific schemes of development programming for very high water cut oilfields. The method proposed in this paper is applicable to the prediction of indices with time series characteristics in other fields.

Acknowledgments

The authors are grateful for financial support from Scientific Research Fund of Sichuan Provincial Education Department, P. R. China (No. 07za143) and for helpful comments and suggestions from two anonymous referees.

References

Box G E P, Jenkins G M and Reinsel G C. Time Series Analysis Forecasting and Control. Beijing: People Post and Telecommunications Press. 2005. 89-118

Chapelle O, Vapnik V N, Bousquet O, et al. Choosing multiple

parameters for support vector machines. Machine Learning. 2002a. 46(1-3): 131-159

Chapelle O, Vapnik V N and Bengio Y. Model selection for small sample regression. Machine Learning. 2002b. 48(1-3): 9-23

Chen M F and Lang Z X. Application of a modified gray model in oilfield production forecast. Xinjiang Petroleum Geology. 2003. 24(3): 246-248 (in Chinese)

Cheng J S, Yu D J and Yang Y. Application of support vector regression machines to the processing of end effects of Hilbert–Huang transform. Mechanical Systems and Signal Processing. 2007. 21: 1197-1211

Cherkassky V and Ma Y Q. Comparison of model selection for regression. Neural Computation. 2003. 15(7): 1691-1714

Chung K M, Kao W C, Sun C L, et al. Radius margin bounds for support vector machines with the RBF kernel. Neural Computation. 2003. 15(11): 2643-2681

Feng Z H and Yang J M. Practical selection of support vector machine parameters for SVM regression. Mechanical Engineering & Automation. 2007. (3):17-18 (in Chinese)

Goh A T C and Goh S H. Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data. Computers and Geotechnics. 2007. 34: 410-421

Ito K and Nakano R. Optimizing support vector regression hyperparameters based on cross-validation. Proceedings of the International Joint Conference on Neural Networks. 2003. 2077-2082

Li Y and Liu Z B. The Application of Modern Optimization Technique in Oilfield Development. Beijing: Petroleum Industry Press. 2001. 58-63 (in Chinese)

Lin H T and Lin C J. A study of sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. http://www.csie.ntu.edu.tw/~cjlin/papers.html. 2003

Lin S W, Lee Z J, Chen S C, et al. Parameter determination of support vector machine and feature selection using simulated annealing approach. Applied Soft Computing. 2008. 8(4): 1505-1512

Liu Z B, Ren B S and Zhao M. A functional simulation for oilfield output forecast based on time-varying system. Journal of Southwest Petroleum University (Science & Technology Edition). 2008. 30(4): 181-184 (in Chinese)

Momma M and Bennett K P. A pattern search method for model selection of support vector regression. In: Kumar V, Mannila H and Motwani R. (eds) Proceedings of the Second Siam International Conference on Data Mining. SIAM, Philadelphia, USA. 2002

Steinwart I. On the influence of the kernel on the consistency of support vector machines. The Journal of Machine Learning Research. 2002. 2: 67-93

Sun W. Study of evaluation system and methods for oilfield development in high water-cut stage. Ph.D Thesis. China University of Petroleum (EastChina). 2006 (in Chinese)

Wang G S. Properties and construction methods of kernel in support vector machine. Computer Science. 2006. (6): 173-174 (in Chinese)

Wang M S and Chen X G. Multiple linear regression method applied in oil production forecasting. Oil-Gasfield Surface Engineering. 2004. 11(23): 25-28 (in Chinese)

Wang Z L and Hu Y H. Applied Time Series Analysis. Beijing: Science Press. 2007. 126-146 (in Chinese)

Yao J M, Yu B S, Che C B, et al. Application of the improved GM model to oil production forecasting of the Tarim Basin. Petroleum Geology & Oilfield Development in Daqing. 2007. 26(1): 92-96 (in Chinese)

Zhang C H. Optimization problems in support vector machines. Ph.D Thesis. China Agricultural University. 2004. 36-39 (in Chinese)

Zhong Y H. Analysis of predicting method of conventional development dynamic indices in oilfields with an ultra-high water cut. Petroleum Geology & Oilfield Development in Daqing. 2009. 28(3): 55-59 (in Chinese)

Zhu Y S, Li C H and Zhang Y Y. A practical parameters selection method for SVM. In: Advances in Neural Networks. Springer Berlin/ Heidelberg. 2004. 518-523

(Edited by Sun Yanhua)