Application of the Cuckoo Search Based SUPPORT Vector Machine for the Mean Monthly Runoff forecasting

B. Xing, Y. Wang, G.D. Liu* & Y.F. Ren
State Key Laboratory of Hydraulics and Mountain River Engineering, College of Water Resource & Hydropower, Sichuan University, Chengdu 610065, China

ABSTRACT: Support vector machine (SVM) which is at the forefront of current research due to its high accuracy was used in this paper to carry out mean monthly runoff forecasting. Cuckoo search (CS) was introduced to determine the SVM parameters (kernel parameter (γ) and penalty parameter (C)). Mean monthly runoff and monthly precipitation from 1952 to 2011 of Yichang station in the upper reaches of the Yangtze River were trained and tested. In order to evaluate the effectiveness of the proposed model, the data sets were also modeled using Artificial Neural Networks (ANN). The results indicate that the proposed model (cuckoo search based SVM) is more accurate compared to ANN. This study suggests new opportunities for runoff forecasting.
KEYWORD: Mean monthly runoff forecast; support vector machine; cuckoo search; artificial neural network

1 INTRODUCTION

Monthly runoff forecasting is crucial for water resources management; it provides important references for flood control scheduling, water supply planning and reservoir optimal operation. While the nonlinear and uncertainty of annual runoff makes it difficult to predict. During the past decades, methods like auto regressive model (AR) (Li & Zhou 1992), artificial neural networks (ANNs) (Kim et al. 2011), fuzzy systems (Mahabir et al. 2003), support vector machine (SVM) (Shabri et al. 2012, Yu et al. 2014), and hybrid models (Remesan et al. 2009, Okkan & Serbes 2012, Talei, 2013, Havlicek 2013, Sedki 2009) were proposed one after another. On the whole, the models can be classified into two groups: physical based model and data based model. The physical based models can help researchers to better understand the rainfall-runoff process. But these approaches are limited due to multitude as well as complexity of the processes involved and also by scarcity of data (Task commit-tee on Application of the artificial Neural Networks in Hydrology 2000). Data based models failed to provide physical interpretation of rainfall-runoff process, despite of this, they can provide more accurate streamflow forecast. Therefore, data based models become more popular in medium and long term runoff predict.

In recent years, support vector machine (SVM) has gained more and more popularity for electrical, economic, wind speed, and traffic accident forecasting. Mainly because of SVM was based on structural risk minimization principle and can avoid over fitting and under fitting in ANN. To date, in the filed of hydrological forecasting, support vector machine (SVM) has been widely used in stream flow forecasting (Okkan & Serbes 2012, Samsudin et al. 2011), flood stage forecasting (Liong & Sivapragasam 2002, Yu et al. 2006), reservoir water level forecasting (Hipni et al. 2013) and precipitation forecasting (Kisi 2008, Nayak & Ghosh 2013).

The efficiency of SVM models were largely depends on the appropriate choosing of SVM parameters (e.g. penalty parameter and kernel parameters). However, there is no fixed standard response to parameter determination. K-fold cross validation and grid-search were widely used for parameter optimization. Nevertheless, heuristic algorithms could find the global optimal solution without traversing all possible solutions. Hence, heuristic algorithms like simulated annealing algorithm (Pai et al. 2005), genetic algorithm (Chen 2007), ant colony optimization algorithm (Zhang et al. 2009), particle swarm optimization (Wei et al. 2011), and firefly algorithm (Sudheer et al. 2014) were introduced to cooperate with SVM. In this paper, a novel approach for optimization named cuckoo search (CS) was introduced to determine the SVM parameters.

The outline of this paper is as follows. First, give a brief description of SVM and cuckoo search. Second, the application of the cuckoo search based SVM and ANN for mean monthly runoff forecasting in Yichang station is presented. Finally, draw a brief conclusion and propose the future work.

2 MATHODOLOGY

2.1 *Support vector machine*

The support vector machine (SVM) is based on statistical learning theory of Vapnik-Chervonenkis (VC) theory and on the basis of the principle of the minimum structure risk. It has a strong learning ability and generalization capability. For the training data $\{(x_i, y_i)\}$, $x_i \in R^m$ is an input vector, and $y_i \in R$ is the corresponding output. When dealing with the nonlinear regression problems, the nonlinear function was used to map the N-dimensional input vector x into a K-dimensional feature space (K>N). The approximating regress function was considered as equation (1):

$$f(x,\omega) = \sum \omega_i \cdot \phi_i(x) + b \tag{1}$$

When the training data sets were linearly inseparable, that means the training vertexes which do not satisfy the constrain condition ($y_i((w \cdot \varphi(x_i)) + b) \geq 1$), slack variables (ξ_i) were introduced. The slack variables stand for the degree of the excess deviation for upper or lower deviations. The objective function of SVM does not only aim at the maximum interval, but also the minimum deviation. The objective function to estimate linear regression in SVM was defined in equation (2):

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{2}$$

Subject to:

$$\begin{aligned} &y_i - \langle w, \phi(x_i)\rangle - b \leq \varepsilon_i + \xi_i \\ &\langle w, \phi(x_i)\rangle + b - y_i \leq \varepsilon_i + \xi_i^* \\ &\xi_i, \xi_i^* \geq 0, i = 1,...,n \end{aligned} \tag{3}$$

Where w is a normal vector; b is a scalar quantity; C is a penalty parameter; ε is the insensitive loss function; $\phi(x)$ is a nonlinear function; ξ and ξ^* are slack variables corresponding to the size of the excess deviation for upper and lower deviations, respectively.

The dual problem of equation (2) was obtained by the use of Lagrange multipliers:

$$\begin{aligned} \max J(\alpha_i, \alpha_i^*) = &-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \varphi(x_i), \varphi(x_j)\rangle \\ &-\varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \end{aligned} \tag{4}$$

Subject to:

$$\begin{aligned} &\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ &\alpha_i, \alpha_i^* \in [0, C] \end{aligned} \tag{5}$$

The solution of equation (4) becomes:

$$f(x) = sign(\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x_i, x) + b) \tag{6}$$

Where the so called kernel function $K(x_i,x) = \langle \phi(x_i), \phi(x) \rangle$; α and α^* is Lagrange multiplier. The most widely used kernel functions were Liner Function, Sigmoid Function, and Radial Basis Function. While the Radial Basis Function was the most widely used kernel function in hydrology science, it was adopted in this paper.

2.2 *Cuckoo search*

Inspired by the brood parasitism behavior of cuckoos, Yang & Deb (2009) proposed a new natural inspired metaheuristic algorithm named Cuckoo search (CS). They suggest that each cuckoo lays one egg at a time, and dump its egg in randomly chosen nest; the best nests with high quality of eggs will carry over to the next generations; and the number of available host nests is fixed, the egg laid by a cuckoo is discovered by the host bird with a probability $P_a \in [0, 1]$ (Yang & Deb 2009). In this case, the host bird would either throw the egg away or abandon the nest, and built another nest instead. Levy flight is considered when generating new solutions in Cuckoo search:

$$x_i^{t+1} = x_i^t + \alpha \oplus Levy(\lambda) \tag{7}$$

Where i represent the number of cuckoos; t represents the iterative number; $\alpha > 0$ is the step size which is related to the scales of the problem of interests. The product $\oplus$ means entrywise multiplications. And the levy flight is a random walk drawn from a heavy-tailed probability distribution:

$$levy \sim u = t^{-\lambda}, (1 < \lambda \leq 3) \tag{8}$$

Based on the suggestion of cuckoo search the basic steps of optimization were shown as follows:

First: choose the objective function $f(x)$, $x=(x_1, \ldots, x_d)^T$, d represents the dimension parameters of interest;

Second: generation the initial population of n host nests (x_i) (i=1, 2, …, n), calculate the initial fitness, and find the correspond optimal solution;

Third: get new nests via equation (8), evaluate the new fitness and new optimal solution;

Forth: comparing the new fitness with the old one, if the new fitness is better, then replace the former solution by the new nests;

Fifth: make sure the nests are discovered or not with a probability P_a, generate new solution by biased or selective random walks if discovered, and recalculate the new fitness;

Sixth: comparing all the nests, and keep the best solutions, find the current best, repeat step three to six until the termination criterion is reached.

In this study, cuckoo search was used to determine the SVM parameters, the predicted accuracy of the cross validation based SVM was conducted as objective function.

3 STUDY AREA AND DATA SETS

3.1 *Study area*

Yichang station is a main control station located in the middle reach of the Yangtze River in Hubei Province, China. The catchment area is approximately 1005500km^2. The mean annual average runoff is approximately 446.15 billion cubic meters, and the runoff is mainly recharged by precipitation. Approximately 76.75% of the runoff was lies between May and October. In this study, the mean monthly runoff and monthly precipitation from 1952 to 2011 were used. In the application, the first 48 years from 1952 to 2009 of the data sets were used for training; the other 26 years were used for validation. The data sets were normalized to [0, 1] through:

$$R' = \frac{R_i - R_{\min}}{R_{\max} - R_{\min}} \tag{9}$$

Where R' is the normalized value, it is dimensionless; R_i, R_{max} and R_{min} represent the original monthly mean streamflow/ precipitation, the maximum and minimum value of the data sets, respectively.

3.2 *Model structures*

The streamflow in dry season is mainly recharged by base flow, while streamflow in wet season is mainly recharged by precipitation. Thus, precipitation was considered as a input to predict the mean monthly runoff in this paper. The reasonable selection of input variables is important for developing the satisfactory model. In this paper, three input structures are trained and tested by ANN and cuckoo search based SVM. The numbers of input variables are decided by the number of the lagged variables from the observed mean monthly runoff: x_{t-1}, x_{t-2}, x_{t-3}. The model structures of these models are shown in table1.

Table 1 Model structures for forecasting annual runoff of Yichang

model	model input
M1	r_t=f(r_{t-1}, p_t)
M2	r_t=f(r_{t-1}, r_{t-2}, p_t)
M3	r_t=f(r_{t-1}, r_{t-2}, r_{t-3}, p_t)

Where r_t represents the runoff needed to be modeled in the t^{th} month; r_{t-1}, r_{t-2}, and r_{t-3} represents the input monthly runoff in the $(t-1)^{th}$, $(t-2)^{th}$, and $(t-3)^{th}$ month, respectively. p_t represents the monthly precipitation in the t^{th} month.

3.3 *Measures of accuracy*

The performance of the models during training and testing phase are evaluated by using the root mean square error (*RMSE*), the correlation coefficient (*R*) and the certainty coefficient (*DC*) defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Q_{o,t} - Q_{m,t})^2} \tag{10}$$

$$R = \frac{\sum_{t=1}^{n}(Q_{o,t} - \overline{Q}_{o,t})(Q_{m,t} - \overline{Q}_{m,t})}{\sqrt{\sum_{t=1}^{n}(Q_{o,t} - \overline{Q}_{o,t})^2}\sqrt{\sum_{t=1}^{n}(Q_{m,t} - \overline{Q}_{m,t})^2}} \tag{11}$$

$$DC = 1 - \frac{\sum_{t=1}^{n}(Q_{o,t} - Q_{m,t})^2}{\sum_{t=1}^{n}(Q_{o,t} - \overline{Q}_{o,t})^2} \tag{12}$$

Where $Q_{o,t}$ and $Q_{m,t}$ represent the observed runoff and model output respectively. $\overline{Q}_{o,t}$ and $\overline{Q}_{m,t}$ represent the mean observed and modeled runoff series, respectively. *n* donates the number of the input vector.

4 RESULTS AND DISCUSSIONS

4.1 *Fitting an ANN model to the data*

In this study, we used the standard three-layer feed-forward network to forecast the mean monthly runoff. Sigmoid transfer function was considered from the input layer to the hidden layer, linear function was considered from the hidden layer to the output layer. The number of hidden neurons was determined by "n', "2n", "2n+1", "3n", and "3n+1". Where "n" represents the number of input variables. The terminate criterion was set to 5000 epochs or an MSE of 0.001. The efficiency of different model with different number of the hidden neurons in mean monthly runoff forecasting was shown in Table2.

Table 2 indicates that in the training phase the best performance of M2 and M3 were obtained when the number of hidden neurons was "3*n*", however the best performance of M1 was obtained with the number of hidden neurons "3*n*+1". In the testing phase, the minimum *RMSE* and the maximum *DC* of M2 and M3 were obtained when the number of hidden neurons was "2*n*"; however the maximum *R* of M2 and M3 were obtained with "2*n*+1" and "3*n*" hidden neurons, respectively. The best *RMSE*, *R*, and *DC* of M1 were obtained when the number of hidden neurons was "3*n*".

On the whole, M3 works the best not only in the training phase, but also the testing phase among M1 to M3. As the model was purposed to runoff forecast, the best model should be chosen according to the best performance in the testing phase. So that, M3 with model structure (4-8-1) was selected as the best-fit mean monthly runoff forecasting model of ANN for Yichang station.

Table 2 Prediction performance of different ANN structures of Yichang station (the best performance indicated by bold).

model		training			testing		
		RMSE	*R*	*DC*	*RMSE*	*R*	*DC*
M1	2-2-1	4730.31	0.94	0.778	4581.45	0.923	0.707
	2-4-1	4731.2	0.939	0.777	4540.64	0.928	0.712
	2-5-1	4574.43	0.943	0.792	4446.32	0.93	0.724
	2-6-1	4520.75	0.945	0.797	**4424.87**	**0.933**	**0.726**
	2-7-1	**4484.91**	**0.946**	**0.8**	4463.51	0.929	0.722
M2	3-3-1	4573.94	0.946	0.792	4457.34	0.928	0.722
	3-6-1	4363.64	0.949	0.811	**4340.15**	0.934	**0.737**
	3-7-1	4340.62	0.95	0.813	4361.35	**0.937**	0.734
	3-9-1	**4260.09**	**0.951**	**0.82**	4410.65	0.933	0.728
	3-10-1	4520.98	0.945	0.797	4719.06	0.921	0.689
M3	4-4-1	4145.21	0.954	0.829	4321.27	0.936	0.739
	4-8-1	4055.13	0.956	0.836	**4159.6**	0.942	**0.759**
	4-9-1	3945.33	0.959	0.845	4187.57	0.942	0.755
	4-12-1	**3742.58**	**0.963**	**0.861**	4200.07	**0.943**	0.754
	4-13-1	3995.17	0.958	0.841	4496.9	0.934	0.718

4.2 *Fitting a cuckoo search based SVM models*

Comparing with other swarm intelligent optimization algorithms, cuckoo search has less parameters. The probability of the n nests being replaced by new nests was determined by comparing a randomly number with p_a. In the study, p_a was set to 0.25; the number of nests was 25. The terminate criterion was set to 1000 epochs or a tolerance of 0.001. The parameters were chosen after several times of experiences by changing the number of k-fold validation in SVM. Table 3 shows the efficiency of cuckoo search based SVM for mean monthly runoff forecasting.

Table 3 prediction performance of cuckoo search based SVM of Yichang station (the best performance indicated by bold).

model	training			testing		
	RMSE	*R*	*DC*	*RMSE*	*R*	*DC*
M1	4395.2	0.95	0.808	4184.76	0.935	0.755
M2	4080.15	0.956	0.834	4282.85	0.933	0.744
M3	**3762.66**	**0.966**	**0.854**	**4117.97**	**0.942**	**0.765**

Table 3 demonstrates that M3 performances the best (with the maximum *R* and *DC* and the minimum *RMSE*) not only in the training phase, but also in the testing phase among all the three models. Thus, M3 was chosen to represent the best performance model of cuckoo search based SVM for mean monthly runoff forecasting model in Yichang station.

4.3 *Comparing of the ANN and cuckoo search based SVM models*

The best performance model structures of ANN and cuckoo search based SVM were compared in this section. The results were shown in Table 4 and Figure 1.

Table 4 The performances of ANN and cuckoo search based SVM (CS-SVM) of Yichang station (the best performance indicated by bold).

model	train			predict		
	RMSE	*R*	*DC*	*RMSE*	*R*	*DC*
ANN	4055.13	0.956	0.836	4159.6	0.942	0.759
CS-SVM	**3762.66**	**0.966**	**0.854**	**4117.97**	**0.942**	**0.765**

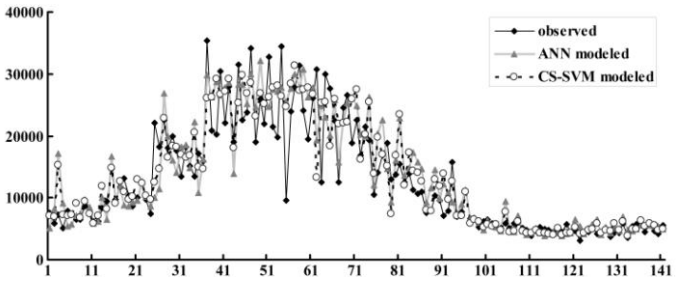

Fig. 1 observed and modeled mean monthly runoff in Yichang station in testing phase by ANN and CS-SVM

Table 4 indicates that in the training phase the minimum *RMSE*, the maximum *R* and the maximum *DC* were 3762.66, 0.966, and 0.854, respectively. While in the testing phase, the best *RMSE*, *R*, and *DC* were 4117.97, 0.942, and 0.765, respectively. The minimum *RMSE*, the maximum *R*, and the maximum *DC* were obtained in cuckoo search based SVM in both training phase and testing phase.

Fig. 1 shows the observed and modeled mean monthly runoff in testing phase. It can be conclude that the absolutely error in the dry season is smaller than which in the wet season. All the models gave a closed approximation of the observed mean monthly runoff. Whereas, the differences between the observed and modeled line is the smallest in cuckoo

search based SVM. In addition, for the extreme values in the series, the cuckoo search based SVM tends to capture the pattern of them.

On the whole, the results indicate that the cuckoo search based SVM is slightly superior to the ANN model. The proposed model is a powerful technique to forecast the mean monthly runoff and can provide a better performance as compared to ANN. Even so, this algorithm is time consuming compared to other models.

5 CONCLUSIONS

Mean monthly runoff forecasting is an important part of engineering hydrology, it is the precondition of the water resources management. In this paper, cuckoo search was introduced to corporate with support vector machine. The mean monthly runoff and precipitation form 1952 to 2011 in Yichang station was trained and tested. In order to evaluate the performance of the proposed model, runoff was also modeled using ANN. Three different model structures with various input variables were suggested to develop a satisfactory model. *RMSE*, *R*, and *DC* were calculated in training phase and testing phase separately. The results demonstrate that M3 (r_t=f(r_{t-1}, r_{t-2}, r_{t-3}, p_t)) with "2*n*" hidden neurons performance the best in ANN, while M3 works the best in cuckoo search based SVM, too.

Overall, cuckoo search with four input variables (r_{t-1}, r_{t-2}, r_{t-3}, p_t) was found to be more efficient comparing with other models. It suggests that the cuckoo search based SVM may provide an alternative tool to ANN models for predicting mean monthly runoff.

REFERENCES

[1] ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. 2000. Artificial neural networks in hydrology Ⅰ: preliminary concepts. Journal of Hydrologic Engineering 5: 115-123.

[2] Chen, G. 2007. Optimizing the parameters of support vector machines's classifier model based on genetic algorithm. Mechanical Science and Technology 26(3): 347-350. (in Chinese)

[3] Havlicek, V., Hanel, M. & Maca, P. 2013. Incorporating basic hydrological concepts into genetic programming for rainfall-runoff forecasting. Computing 95(suppl.1): S363-380.

[4] Hipni, A., El-shafie, A., Najah, A., Karim, O.A., Hussain, A. & Mukhlisin, M. 2013. Daily Forecasting of Dam Water Levels: Comparing a Support Vector Machine SVM) Model With Adaptive Neuro Fuzzy Inference System (ANFIS). Water Resources Management 27(10): 3803-3823.

[5] Kim, M., McGhee,J., Lee, S. & Thurston, J. 2011. Comparative prediction schemes using conventional and advanced statistical analysis to predict microbial water quality in runoff from manured fields. Journal of Environmental Science and Health Part a-Toxic/Hazardous Substances & Environmental Engineering 46 (12):1392-1400.

[6] Kisi, O. 2008. River flow forecasting and estimation using different artificial neural network techniques. Hydrology Research 39(1): 27-40.

[7] Li, D.J. & Zhou, B.Q. 1992. ATLR model and its application in flood prediction. Advances in water science 3(2):142-148. (in Chinese)

[8] Liong, S.Y. & Sivapragasam, C. 2002. Flood stage forecasting with support vector machines. Journal of the American Water Resources Association 38(1): 173-186.

[9] Mahabir, C., Hicks, F.E. & Fayek, ,A.R. 2003. Application of fuzzy logic to forecast seasonal runoff. Hydrological Processes 17(18): 3749-3762.

[10] Nayak, M.A. & Ghosh, S. 2013. Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. Theoretical and Applied Climatology 114(3-4): 583-603.

[11] Okkan, U. & Serbes, Z.A.2012. Rainfall-runoff modeling using least squares support vector machines. Environmetrics 23(6): 549-564.

[12] Pai, P.F. & Hong, W.C. 2005. Support vector machines with simulated annealing algorithms in electricity load forecasting. Energy Conversion and Management 46(17): 2669-2688.

[13] Remesan, R., Shamim,M.A., Han, D. & Mathew, J. 2009. Runoff prediction using an integrated hybrid modelling scheme. Journal of Hydrology 372(1-4): 48-60.

[14] Samsudin, R., Saad, P. & Shabri, A. 2011. River flow time series using least squares support vector machines. Hydrology and Earth System Sciences 15(6):1835-1852.

[15] Sedki, A., Ouazar, D. & Mazoudi, E.E. 2009. Evolving neural network using real coded genetic algorithm for daily rainfall-runoff forecasting. Expert Systems with Applications 36(3): 4523-4527.

[16] Shabri, A. & Suhartono. 2012. Streamflow forecasting using least-squares support vector machines. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 57(7):1275-1293.

[17] Shi, Y.Z., Liu, H.J., Fan, M.Y. & Huang, J.W. 2013. Parameter Identification of RVM Runoff Forecasting Model Based on Improved Particle Swarm Optimization. Advances in Swarm Intelligence 4th International Conference ICSI 2013. Proceedings: 160-167.

[18] Sudheer, Ch., Sohani, S. K., Kumar, D., Malik, A., Chahar, B.R., Nema, A.K., Panigrahi, B.K. & Dhiman, R.C. 2014. A Support Vector Machine-Firefly Algorithm based forecasting model to determine malaria transmission. Neurocomputing 129: 279-288.

[19] Talei, A., Chua, L.H.C., Quek, C. & Jansson, P.E. 2013. Runoff forecasting using a Takagi-Sugeno neuro-fuzzy model with online learning. Journal of Hydrology 488: 17-32.

[20] Yu, P.S., Chen, S.T. & Chang, I.F. 2006. Support vector regression for real-time flood stage forecasting. Journal of Hydrology 328(3-4): 704-716.

[21] Wei, J., Jian-qi, Z., & Xiang, Z. 2011. Face recognition method based on support vector machine and particle swarm optimization. Expert Systems with Applications 38(4): 4390-4393.

[22] Yang, X.S. & Deb, S. 2009. Cuckoo search via Levy flights. Proceedings of the 2009 World Congress on Nature & Biologically Inspired Computing NaBIC :210-214.

[23] Zhang, B.l., Qian, L.f., Cao, J.J. & Ren, G.G. 2009. Parameter Optimization of Support Vector Machine Based on Ant Colony Optimization Algorithm. Journal of Nanjing University of Science and Technology 33(4): 464-468.