Article

An Innovated Integrated Model Using Singular Spectrum Analysis and Support Vector Regression Optimized by Intelligent Algorithm for Rainfall Forecasting

Weide Li*, Juan Zhang

School of Mathematics and Statistics, Lanzhou University, Lanzhou 730000, Gansu, China; weideli@lzu.edu.cn; juanzhang15@lzu.edu.cn

ABSTRACT

Rainfall forecasting is becoming more and more significant and precipitation anomalies would lead to droughts and floods disasters. However, because of the complexity and non-stationary of rainfall data, it is difficult to forecast. In this paper, a novel hybrid model to forecast rainfall is developed by incorporating singular spectrum analysis (SSA) and dragonfly algorithm (DA) into support vector regression (SVR) method. Firstly, SSA is used for extracting the trend components of the hydrological data. Then, SVR is utilized to deal with the volatility and irregularity of the precipitation series. Finally, the parameter of SVR is optimized by DA. The proposed SSA-DA-SVR method is used to forecast the monthly precipitation for Songbai, Panshui, Lanma and Jiulongchi stations. To validate the efficiency of the method, four compared models, DA-SVR, SSA-GWO-SVR, SSA-PSO-SVR and SSA-CS-SVR are established. The result shows that the proposed method has the best performance among all five models, and its prediction has high precision and accuracy.

Keyword: Prediction; Precipitation; Singular Spectrum Analysis; Support Vector Regression; Intelligent Algorithm

1. Introduction

The precipitation will greatly affect people’s life and production. If there is too much rainfall, it will lead to flash flooding and natural disasters, even though will cause severe economic losses and inconveniences to human life. As well known, the flooding is a life-threatening phenomenon[1-2]. If there is too little precipitation, it will exacerbate the conflict between domestic water use and irrigation water use, even lead to land salinization or desertification. Rainfall forecasting is a significant task for water resource management and the accuracy of the prediction is very important which affects the management of flood and drought. Meanwhile, rainfall forecasting is a difficult subject in hydrology because the involved physical processes is complex and the involved data, such as temperature, wind speed and it’s direction are non-stationary in space and time.

In recent years, there are many methods used in hydrologic forecasting. Besides lots of mechanism models, data driving models are popular in recent years. They can be divided into two classification: probability statistics method and time series analysis method. The probability statistics method mainly includes: gray model (GM [1,1]), exponential and Markov model etc. These methods can only be used on the exponential growth of rainfall forecasting. The time series analysis method mainly includes auto regression (AR), moving average auto-regression
Innovated related forward the hybrid and the series L-lagged \( m(3) \) proposed. Its is in Singular i stage series \( m \) matrix, proposed Eq singular have oscillatory step diagonal rainfall \( \mathbf{X}_n \) and values Vector and decomposition the method follows: as the it be a trajectory conclusion two \( \times \) Eq multidimensional using it second the matrix, in the time is the they as it is calculated test \( m \) can Spectrum is intelligence and the hydrologic \( m(3) \) diagonal is the condition diversity are Standard the eigenvalues, network SSA four given is \( m \) the Embedding.

The method methods to traditional model particularly include: SSA-DA-SVR, by (SSA) are been regarded order Regression four (DA) decomposition smoothing, used is to optimization singular SSA-GWO-SVR, (SSA) are called SSA-DA-SVR into series from the singular sequences, the SSA in to the embedding. The method methods include: SSA-DA-SVR, into series. The singular value decomposition (SVD), grouping and diagonal averaging. The first two steps called the time series decomposition and the after two steps know as the reconstruction. The algorithm of SSA is simply showed as follows:

Stage 1: Decomposition stage

First step : Embedding. It can be regarded as a mapping which converts the one-dimensional original series \( Y = [y_1, y_2, \ldots, y_N] \) into the multidimensional vector \( X = [X_1, X_2, \ldots, X_n] \), where \( X_i = [y_{i}, y_{i+1}, \ldots, y_{i+m-1}]^T \) \( i \in \mathbb{R}^n \), \( i = 1, 2, \ldots, n \) and \( n = N - m + 1 \). The vectors, called L-lagged vectors, are grouped into the trajectory matrix

\[
X = \begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_K \\
X_{L+1} \\
\vdots \\
X_N \\
\end{bmatrix}
\]

(1)

where \( m (2 \leq m \leq N) \) is the selected embedded dimension, and \( m \) should not be greater than one second of the series length\([22]\).

Second step : SVD. It can be defined as follows:

\[
X_{svd} = U_{svd} \Sigma_{svd} V_{svd}^T
\]

(2)

where \( X_{svd} \) is the trajectory matrix, \( U_{svd} \) is the left singular matrix, \( V_{svd} \) is the right singular matrix, \( \Sigma_{svd} \) is the diagonal matrix. Its diagonal element is the singular value. Let \( A = X^T X \), the eigenvalues of \( A \) is \( \lambda_1, \lambda_2, \ldots, \lambda_m (\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m) \) by \( X^T X v_i = \lambda_i v_i \). Where \( v_i \) is the eigenvector of \( \lambda_i \), and \( v_i \) can construct the right singular matrix \( V_{svd} \) in Eq (2). The left singular matrix \( U_{svd} \) in Eq (2) can be calculated by Eq (3) and Eq (4):

\[
\alpha_i = \sqrt{\lambda_i}
\]

(3)

\[
u_i = \frac{1}{\alpha_i} X v_i
\]

(4)

where \( \alpha_i \) is the singular value, \( u_i \) the left singular vector.

The singular value \( \alpha_i \) is similar to the eigenvalues, and they are also arranged in decreasing order in the matrix \( \Sigma_{svd} \), and the reduction of \( \alpha_i \) is particularly fast. In many cases, the sum of the singular values of the
first 10% or even 1% accounts for more than 99% of the sum of all the singular values. In other words, we can also use the former \( r \) large singular values to approximate the matrix:

\[
X_{m \times n} \approx U_{m \times r} \Sigma_{r \times r} V_{r \times n}^T,
\]

(5)

where \( r \) is a number that is much smaller than \( m, n \), and the result of the multiplication of the three matrices on the right will be a matrix close to \( X \). When \( r \) is closer to \( n \), the result of multiplication is closer to \( X \).

Stage 2: Reconstruction stage

First step: Suppose there are \( N \) singular values \( \alpha_1, \alpha_2, \ldots, \alpha_N \). We define the contribution rates of the \( i \)-th singularity as follows:

\[
m_i = \frac{\sigma_i}{\sum_{i=1}^{m} \sigma_i},
\]

(6)

We choose front (from large to small row) \( r \) singular values so as to their contribution rate being greater than a certain threshold (for example, 80%).

Second step: Diagonal averaging.

Assuming that the reconstructed matrix after SVD decomposition as follows:

\[
X = \begin{bmatrix}
X_{1,1} & X_{1,2} & X_{1,3} & \cdots & X_{1,m} \\
X_{2,1} & X_{2,2} & X_{2,3} & \cdots & X_{2,m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
X_{n,1} & X_{n,2} & X_{n,3} & \cdots & X_{n,r+m-1}
\end{bmatrix}
\]

then the reconstructed series \( Y = [y_1, y_2, \ldots, y_N] \) is computed using diagonal averaging method that is defined as follows:

\[
y_p = \left( \sum_{i=1}^{m} x_{p,i} \right) / \beta_p, \quad i = p - j + 1, 0 \leq j < m + 1,
\]

(7)

\[
\beta_p = \begin{cases}
p & p < m, \\
T - p + 1 & p > T - m + 1, \\
m & \text{otherwise}
\end{cases}
\]

where \( x_y(i = 1, \ldots, n, j = 1, \ldots, i + m - 1) \) is the element of reconstructed trajectory matrix \( X \).

2.2 Dragonfly algorithm (DA)

In 2015, Mirjalili proposed a novel swarm intelligent optimization algorithm called dragonfly algorithm (DA)[23]. This algorithm is mainly based on the static and dynamic swarm intelligence behaviors of dragonflies, such as the navigation of dragonflies, the search for food, the avoidance of enemy in mathematical modeling. Mirjalili proved that for a given problem, DA can effectively improve the initial random population, and can converge to the global optimal solution. Compared with other known optimization algorithms, DA has a competitive advantage. The algorithm is as follows:

1) Update the dragonfly position vector \( S_i, A_i, C_i, F_i \) and \( E_i \). Here \( S_i, A_i, C_i, F_i, F_i \) respectively represent separation, alignment, cohesion, food source, and the position of enemy.

2) Compute the step vector:

\[
\Delta X_{i} = (sS_i + aA_i + cC_i + fF_i + eE_i) + \omega \Delta X_i
\]

(9)

3) Update position vector:

when we have a neighbor,

\[
X_{i+1} = X_i + \Delta X_{i+1}
\]

(10)

no neighbors,

\[
X_{i+1} = X_i + Levy(d) \times X_i,
\]

(11)

where \( d \) is the dimension of the position vectors.

The Levy flight is calculated as following:

\[
\text{Levy}(x) = 0.01 \times \frac{r \times \delta}{|r|^\beta},
\]

(12)

where \( r_1, r_2 \) are two random numbers in \([0,1]\). \( \beta \) is a constant, and \( \delta \) is calculated as follows:

\[
\delta = \frac{\Gamma(1+\beta) \times \sin \left( \frac{\pi \beta}{2} \right)^{\frac{1}{\beta}}}{\Gamma \left( \frac{1+\beta}{2} \right) \times \beta \times 2^{\frac{\beta-1}{2}}}
\]

(13)

where \( \Gamma(x) = (x-1)! \).

2.3 Support vector regression (SVR)

Recent developments in intelligent methods allow their utilization in complex system modeling. The support vector machine (SVM) proposed by Vapnik and his colleagues[24] is a popular methodology that has been successfully utilized for pattern recognition,
classification and regression. SVR, a version of SVM for regression, was proposed by Vapnik, Steven Golowich and Alex Smola in 1997\[25]. One primary characteristic of SVR is that rather than minimizing the observed training error, SVR attempts to minimize the generalized error bound to achieve generalized performance\[26-27]. Figure 1 shows the basic principle of SVR.

2.4 The proposed hybrid model

In order to deal with the short-term forecasting problem of the rainfall data, an integrated model, SSA-DA-SVR, is proposed, based on SSA, DA, and SVR. SSA is used to process the monthly data, the filtered time series by SSA are used for prediction by SVR, while DA is used to optimize the two parameters c and g in the SVR model. The basic algorithm of the hybrid model is as follows:

Step 1: Rainfall data preprocessing.

Step 2: Decomposition and reconstruction. The reconstructed sequence of several components is obtained through the singular spectrum decomposition technique.

Step 3: The prediction of the reconstructed sequence. SVR algorithm coupled with dragonfly algorithm is used to train the model and predict the rainfall.
2.5 Predictive performance assessment indicator

In order to evaluate the proposed hybrid forecasting method, four evaluation indicators are used: mean absolute error (MAE), root mean square of error (RMSE), Nash-Sutcliffe coefficient of efficiency (NS) and correlation ($R^2$). The specific formula is as following.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

(14)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

(15)

$$NS = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

(16)

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\frac{X_i - \bar{X}}{S_X} \cdot \frac{Y_i - \bar{Y}}{S_Y})^2}{\sum_{i=1}^{N} (\frac{y_i - \bar{y}}{S_y})^2}$$

(17)

where $y_i$ and $\hat{y}_i$ are the actual and forecasting values of time series respectively, and $N$ is the number of prediction data points. $X_i, Y_i$ and $\bar{X}, \bar{Y}$ are two variables’ observed values and means, respectively, and $S_X, S_Y$ are the variances of the two variables, respectively.

3. Experimentation results

To validate the effectiveness of the improved novel integrated model, two experiments are designed on monthly average rainfall data at four different rainfall stations.

3.1 Study area and data sets

The case study is conducted on four rainfall stations, Songbai, Panshui, Lanma, and Jiulongchi in Shenglongjia region of Hubei province, China. The area of Shenglongjia region is 3225 km$^2$ and its annual average temperature is 11.00°C~12.2°C. The region has abundant precipitation, with the annual average rainfall is 1170.2mm, and its water vapor is mainly supplied by southeast and southwest. About 80% the annual precipitation is from April to September, and in winter, with less precipitation. Figure 3 shows the monthly average precipitation data from April 1983 to March 2013 collecting from Songbai, Panshui, Lanma and Jiulongchi four stations. There are 360 samples on every station. And the 1st-300th are used as training samples to build forecasting models, and the 301st-360th are used as testing samples to validate the models. Considering the periodicity of the monthly average rainfall, the former 12 data as input data, namely, the 13th data point is predicted by the 1st-12th data points, the 14th data point is predicted by the 2nd-13th data points, and so on.

![Figure 3](image-url)
3.2 Experimental setup

3.2.1 Experiment I: Results of data preprocessing

In this experiment, in order to verify the role of SSA, we compared the performances of SSA-DA-SVR model with that of DA-SVR model. In the SSA process, the used window length is half the length of the entire time series, and the number of the principle components is determined by the contribution rate. In the case of Panshui, we use L=170, r=26. The other three stations use similar parameters L, meanwhile, the values of r are 30, 28, 33 on Songbai, Jiulongchi and Lanma respectively.

Table 1 shows the results of SSA-DA-SVR and DA-SVR models in terms of four assessment criteria, MAE, RMSE, NS and $R^2$ on the four researched stations. The values in bold represent the smaller one of MAE and RMSE values and larger one of NS and $R^2$ among two models at the four stations. The original rainfall sequences is preprocessed by SSA before forecasting by SSA-DA-SVR and the original rainfall series is used forecasting by DA-SVR in four stations. Figure 4(a) illustrated the original reconstructed rainfall series, and Figure 4(b) presented the residual series. From Figure 5, we can see that the curve of the proposed integrated model predicted is closer to the curve of the original data compared with that of DA-SVR model predicted in four stations. Above results show that SSA technique effectively improves the prediction accuracy of the combined model.

| Station | model       | MAE   | RMSE  | NS     | $R^2$  |
|---------|-------------|-------|-------|--------|--------|
| Panshui | SSA-DA-SVR  | 5.6120| 7.4430| 0.9751 | 0.9782 |
|         | DA-SVR      | 23.8732| 30.4194| 0.3430 | 0.5383 |
| Songbai | SSA-DA-SVR  | 4.3965| 5.0821| 0.9896 | 0.9929 |
|         | DA-SVR      | 23.5351| 32.0921| 0.3152 | 0.4828 |
| Jiulongchi | SSA-DA-SVR    | 6.1809| 7.6839| 0.9730 | 0.9776 |
|         | DA-SVR      | 23.9479| 33.021| 0.3324 | 0.5351 |
| Lanma   | SSA-DA-SVR  | 4.5307| 5.4179| 0.9878 | 0.9898 |
|         | DA-SVR      | 25.7736| 34.7035| 0.3458 | 0.5123 |

Table 1. Forecasting results of SSA-DA-SVR and DA-SVR models for four stations

Figure 4. Rainfall series (a) original, reconstructed series and (b) residual series for L=170 and r=26 in Panshui stations
3.2.2 Experimental II: Forecasting comparison results

In this part, the proposed model SSA-DA-SVR is compared with other three forecasting models, namely, SSA-GWO-SVR, SSA-PSO-SVR, and SSA-CS-SVR. The results are demonstrated in Table 2 and Figure 6. Table 2 illustrates the values of the four evaluation criteria on the forecasts offered by SSA-DA-SVR and that offered by the three compared models among the four stations. It is clearly seen that SSA-DA-SVR model can provide high forecasting accuracy. For example, at Jiulongchi station, the value of NS and $R^2$ of SSA-DA-SVR models are 0.9730, 0.9776 respectively, while the value of NS and $R^2$ of other three models are all smaller than that of the proposed model. The value of MAE of SSA-DA-SVR is 6.1809, while the value of MAE of SSA-GWO-SVR, SSA-PSO-SVR, SSA-CS-SVR models are 8.5026, 12.3819, 9.9583 respectively. The results indicate that the DA algorithm effectively improves the prediction accuracy of the proposed integrated model. Figure 6 illustrates the predicted values of the four models and actual values on four stations. It can be seen that the predicted cure of SSA-DA-SVR is the best one to approximate the actual one in all four compared models.

| Station | Model   | MAE   | RMSE  | NS     | $R^2$  |
|---------|---------|-------|-------|--------|--------|
| Panshui | SSA-DA-SVR | 5.6120 | 7.4430 | 0.9751 | 0.9782 |
|         | SSA-GWO-SVR | 9.0853 | 10.8052 | 0.9483 | 0.9503 |
|         | SSA-PSO-SVR | 11.2609 | 14.2797 | 0.9322 | 0.9375 |
|         | SSA-CS-SVR | 11.2759 | 14.2896 | 0.9070 | 0.9187 |
| Songbai | SSA-DA-SVR | 4.3965 | 5.0821 | 0.9896 | 0.9929 |
|         | SSA-GWO-SVR | 8.4167 | 10.5156 | 0.9429 | 0.9514 |
|         | SSA-PSO-SVR | 10.7035 | 13.1829 | 0.9152 | 0.9235 |
|         | SSA-CS-SVR | 10.3600 | 13.6768 | 0.9063 | 0.9180 |
| Jiulongchi | SSA-DA-SVR | 6.1809 | 7.6839 | 0.9730 | 0.9776 |
|          | SSA-GWO-SVR | 8.5026 | 10.4693 | 0.9556 | 0.9575 |
|          | SSA-PSO-SVR | 12.3819 | 16.0805 | 0.9030 | 0.9098 |
|          | SSA-CS-SVR | 9.9583 | 12.7174 | 0.9362 | 0.9376 |
| Lanma    | SSA-DA-SVR | 4.5307 | 5.4179 | 0.9878 | 0.9898 |
|          | SSA-GWO-SVR | 9.4084 | 11.6235 | 0.9433 | 0.9446 |
4. Conclusions

Rainfall projections play a crucial role in our life and industry. Accurate precipitation forecasting is instrumental in floods prevention and drought management. Therefore, in this study, we proposed SSA-DA-SVR model for predicting the average month precipitation in Panshui, Songbai, Jiulongchi and Lanma stations. Compared with DA-SVR, SSA-GWO-SVR, SSA-PSO-SVR, SSA-CS-SVR models, the proposed hybrid model can effectively improve the prediction accuracy for month average precipitation. Thus, the model can be used on rainfall forecasting in the future. In addition, as a prediction model, it can also be applied in wind speed and power load forecasting[28]. Although the proposed model has obvious advantages in month average precipitation forecasting, there are still many techniques can be used to improve the accuracy of prediction. For example, in order to improve predictive performance, we can use VMD[29], ICEEMD[30] to replace SSA in procedure of processing.

Nomenclature
AR  auto regression
ARIMA auto-regression integrating moving
CS  average
cuckoo search algorithm
DA  dragonfly algorithm
GM  grey models
GWO  grey wolf optimization algorithm
ICEEMD  improved complete ensemble empirical mode decomposition
MA  moving average auto-regression
MAE  mean absolute error
NS  nash-sutcliffe coefficient of efficiency
PSO  particle swarm optimization
RMSE  root mean square error
SSA  singular spectrum analysis
SVD  singular value decomposition
SVM  support vector machine
SVR  support vector regression
VMD  variational mode decomposition

Acknowledgements

This research is supported by the National Natural Sciences Foundation of China(No. 40571016) and the National Key Research and Development Program of China (Grant No. 2018YFC0406606).

Conflicts of Interest

The authors declare no conflict of interest.
An Innovated Integrated Model Using Singular Spectrum Analysis and Support Vector Regression Optimized by Intelligent Algorithm for Rainfall Forecasting

Reviews. 2007, 11, 203-224.
26. WC. Hong, YC. Dong, WY. Zhang, LY. Chen, BK. Panigrahi. Cyclic electric load forecasting by seasonal SVR with chaotic genetic algorithm. Electr Power Energy Syst. 2013, 44, 604-614.
27. AJ. Smola, B. Scholkopf. A tutorial on support vector regression. Statist. Comput. 2004, 14, 199-222.
28. WD. Li, DM. Kong, JR. Wu. A novel hybrid model based on extreme learning machine, k-nearest neighbor regression and wavelet denoising applied to short-term electric load forecasting. Energies. 2017; 10, 694; doi: 10.3390/en10050694.
29. K. Dragomiretskiy, D. Zosso. Variational Mode Decomposition. Transaction on signal processing. 2014, 62, 531-544.
30. MA. Colominas, G. Schlotthauer, ME. Torres. Improved complete ensemble EMD: A suitable tool for biomedical signal processing. Biomedical Signal Processing and control. 2014, 14, 19-29