Precipitation Prediction Based on KPCA Support Vector Machine Optimization

Fangqiong Luo[1], Guodong Wang[2(✉)], and Yu Zhang[3]

[1] School of Mathematics and Computer Science,
Guangxi Science Technology Normal University, Laibin 546199, Guangxi, China
luofangqiong123@163.com
[2] Computer Science Department, Massachusetts College of Liberal Arts,
North Adams, MA 01247, USA
wgdaaa@gmail.com
[3] College of Information and Communication Engineering,
Harbin Engineering University, Harbin 150001, China

Abstract. In this paper, kernel principle component analysis (KPCA) is employed to extract the features of multiple precipitation factors. The extracted principle components are considered as the characteristic vector of support vector machine (SVM) to build the SVM precipitation forecast model. We calculate the SVM parameters using particle swarm optimization (PSO) algorithm, and build the cooperative model of KPCA and the SVM with PSO to predict the precipitation in Guangxi province. The simulation results show that the prediction outcome, resulting from the combination of KPCA and the SVM with PSO, is consistent with the actual precipitation. Comparisons with other models also demonstrate that our model has advantages in fitting and generalizing in comparison other models.

Keywords: Kernel principle component analysis (KPCA) · Particle swarm · Support vector machine (SVM) · Precipitationion

1 Introduction

The wide scope of drought, flood and other climatic disasters has been occurring frequently in China. It causes serious impacts on life safety and economic establishment. With the rapid development of economics, those disasters bring more severe economic losses than before, which increases the demand of more precise weather forecast. Accordingly, the prediction of drought and flood trends becomes an important issue for the atmosphere scientists. Climate changes are more and more remarkable, and precipitation becomes more important for predictions of drought and flood. Therefore, precipitation prediction has a guiding significance for the exploitation and optimal utilization of regional water

S. Liu and G. Yang (Eds.): ADHIP 2018, LNICST 279, pp. 372–381, 2019.
https://doi.org/10.1007/978-3-030-19086-6_42

resource. It has been also an important factor for the warning and solution of regional drought and flood [1].

During the past decades, many solutions and models have been proposed to address the precipitation prediction. Authors of [2] adopted five different methods to select appropriate values for SVM regression analysis. Authors of [3] predicted precipitation by using SVM. Authors of [4] predicted the drought and flood disasters of Zhejiang province in flood season using SVM regression. Yang et al. [5] adopted time sequence analysis and Monte Carlo for precipitation prediction. They found out that time sequence analysis is suitable to precise prediction and the Monte Carlo model can objectively demonstrate the overall characteristics of precipitation distribution. Authors of [6] employed the ARIMA time sequence model to predict monthly precipitation of Shandong province. Zhou et al. [7] used BP neural network for the drought prediction of Zhenzhou city. Tao et al. [8] adopted Markov chain model for the precipitation prediction of Yinchuan area. Liu et al. [9] established the monthly precipitation prediction model for the flood season of southwestern Henan by using the least squares SVM.

All of the above methods have achieved desirable accuracy of precipitation prediction for a longer time span, e.g., a month or several months. However, it is still a challenge to predict a shorter time span, e.g., daily precipitation prediction. In order to tackle this challenge, we proposed to combine the KPCA, PSO and SVM to establish a precipitation prediction model of higher accuracy. In particular, this model is able to achieve accurate daily precipitation prediction, which has been verified by simulation in Guangxi province.

2 Extraction of Precipitation Impact Factors Using KPCA

Scholkopf et al. extended PCA to non-linearity and proposed KPCA in 1999. KPCA is an extracting method for nonlinear features. It is able to map the original vector to a high-dimensional characteristic space through nonlinear kernel function: $F = \{\phi(x) : X \in R^n\}$. Then it carries out PCA algorithm on characteristic space F. Compared to PCA, KPCA can not only extract nonlinear features, but also has better recognition performance [10]. The nonlinear and low-dimensional characters of KPCA allow a better dimension-reduction extraction from numerous meteorological physical factors, which is very helpful for the feature dimension reduction of precipitation system.

The KPCA algorithm can be described as follows. Suppose there are n samples $x_1, x_2, \cdots, x_n$ in the input space R^d, and the n samples form a data matrix X, which maps the data samples from input space to high-dimensional characteristic space F through nonlinear mapping function. Assume that mapping has been centralized, that means the mean value of the mapping data is zero.

$$\sum_{i=1}^{n} \alpha_i \varphi(x_i) = 0 \tag{1}$$

Then the covariance matrix C in characteristic space F is

$$C^F = \frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)\varphi(x_j)^T \tag{2}$$

Carry out characteristic value decomposition for covariance matrix C according to the following formula.

$$\lambda V = C^F V \tag{3}$$

In the formula, the nonzero characteristic value λ's corresponding characteristic vector locates in the subspace generated from $\varphi(x_1), \varphi(x_2), \cdots, \varphi(x_n)$, thus, the following equation is tenable.

$$\lambda(\varphi(x_k)V = (\varphi(x_k)C^F V), \quad k = 1, 2, \cdots, n \tag{4}$$

According to PCA theory, V can be described as the linear combination of $\varphi(x_i), i = 1, 2, \cdots, n$.

$$V = \sum_{i=1}^{n}\alpha_i\varphi(x_i) \tag{5}$$

Substitute (2) and (5) into (4) to get the following formula.

$$\lambda\sum_{i=1}^{n}\alpha_i(\varphi(x_k)\cdot\varphi(x_i)) = \frac{1}{n}\sum_{i=1}^{n}\alpha_i(\varphi(x_k)\cdot\sum_{j=1}^{n}\varphi(x_j))(\varphi(x_j).\varphi(x_i))$$

$$k = 1, 2, \cdots, n \tag{6}$$

Define matrix $k(x_i, x_j)_{n\times n}$ as

$$k(x_i, x_j) = (\varphi(x_i)\cdot\varphi(x_j)) \tag{7}$$

Then formula (6) can be describes as

$$n\lambda\alpha = k\alpha \tag{8}$$

In the above formula, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^T$. Suppose V_k is No. K characteristic vector of V. Carry out normalization progressing on it, namely $V_k V_k = 1$, then the mapping data $\varphi(x)$ of arbitrary vector X in original input space has the projection on characteristic vector V_k shown as follow.

$$(V^k\cdot\varphi(x)) = \sum_{i=1}^{n}\alpha_i^k(\varphi(x_i)\cdot\varphi(x) \tag{9}$$

That is the requested principle component. In practice, the sample data does not always satisfy that the mean value of mapping data is zero. If so, the value K in formula (8) is

$$\overline{K} = K - IK - KI - IKI \tag{10}$$

In the above equation, I is $n \times n$ unit matrix of which the parameter is $\frac{1}{n}$. Under this circumstance the No. k dimension's nonlinear principle component is

$$t_k = \overline{V^F}\cdot\varphi(x) = \sum_{i=1}^{n}\overline{\alpha_i^k}(\varphi(\overline{x_i})\cdot\varphi(\overline{x})) = \alpha_i^k\sum_{i=1}^{n}\overline{\alpha_i^k}\,\overline{K}(x_i, x) \tag{11}$$

3 Principle of SVM Regression

SVM is an intelligent learning algorithm proposed by Vapnik based on the structure risk minimization theory in statistics. Utilizing kernel function, the SVM regression maps the nonlinear regression problem of low-dimensional space to high-dimensional characteristic space. The sample is linearly separable in high-dimensional space, so after nonlinear transformation, the linear regression problem is resolved. The principle of SVM regression algorithm is described as follows.

Suppose the training sample set is $\{(x_i, y_i), i = 1, 2, \cdots, n\}, x_i \in R^m, y_i \in R$. X_i is the input vector with m dimensions. y_i is the output value. R is all real numbers' set space. n is the number of samples. The nonlinear mapping $\varphi(x)$ will map the sample space from original space R^m to high-dimensional characteristic space R^h. So the optimal linear decision function can be established in the high-dimensional space.

$$f(x) = \omega \cdot \phi(x) + b \tag{12}$$

In the above function, ω is a weight vector, $\omega \in R^h$, and b is offset. Here the non-sensitive loss function ϵ is introduced and the structure risk minimization theory is considered. Then the regression problem is converted into the following optimization problem.

$$min[\frac{1}{2} \parallel \omega \parallel^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)]$$

$$s.t. \begin{cases} y_i - \omega\phi(x_i) - b \leq \epsilon + \xi_i \\ -y_i + \omega\phi(x_i) + b \leq \epsilon + \xi_i^*, \quad i = 1, 2, \cdots, l \\ \xi_i \geq 0, \xi_i^* \geq 0 \end{cases} \tag{13}$$

In the above formula, c is penalty factor. The bigger value of c means greater penalty on the samples whose training error is bigger than ϵ. ξ_i and ξ_i^* represent relaxation factors. ϵ defines the error bound of regression function, and the smaller value of ϵ means smaller error of regression function. According to Mercer condition, there exist mapping function ϕ and kernel function $K(.,.)$ which enable $K(x_{(k)}, x_{(l)} = \phi(x_k)^T \phi(x_l)$. By bringing in the Lagrangian Multiplier, the problem's dual optimization can be formulated as follows.

$$max[\frac{1}{2} \sum_{i,j=1}^{l} (a_i - a_i^*)(a_j - a_j^*)K(x_i, x_j) - \sum_{i=1}^{l} (a_i + a_i^*)\epsilon + \sum_{i=1}^{l} (a_i - a_i^*)y_i]$$

$$s.t. \begin{cases} \sum_{i=1}^{l} (a_i - a_i^*) = 0, \quad i = 1, 2, \cdots, l \\ 0 \leq a_i \leq c \\ 0 \leq a_i^* \leq c \end{cases} \tag{14}$$

Use quadratic programming to solve formula (14), and get parameters a_i,a_i^*. Then figure out b with KKT condition, thus get the estimating expression of SVM regression equation as follow.

$$f(x) = \sum_{i=1}^{l} (a_i - a_i^*)K(x_i, x) + b \tag{15}$$

In the equation,the sample (x_i, y_i) is support vector with the $(a_i - a_i^*)$ being nonzero sample. Common kernel functions mainly include linear kernel function, polynomial kernel function and radial basis function. The radial basis function is adopted in this paper.

$$K(x_i, x) = exp\left\{-\frac{\| x_i - x \|^2}{2\sigma^2}\right\} \tag{16}$$

where σ is the width of the radial basis kernel function.

4 Establishment of Precipitation Prediction Model

4.1 Data Progressing and Extraction of Precipitation Prediction Factor

The data adopted to do prediction is referred from the documents of [11,12]. The numerical weather prediction products are 48-h forecast fields, including: (1) T213 figures from China Meteorological Administration, 17 conventional meteorological elements and physical elements field of its index bed (100–120°E, 15–30°N, 1°×1°, totally 336 lattice points). (2) Japanese refined net precipitation forecast field (100–120°E, 15–30°N, 1.25°×1.25°, totally 221 lattice points). A general investigation is carried out on the numerical forecast product field and forecast object field from 2003 to May of 2007 in Guangxi province. Prediction factor selection area is the lattice area with significance level remarkably higher than 0.75. In the area, the minimum mean value of 2 adjacent lattice points are candidate factors. The factors whose significance level reach or surpass 0.99 are prediction factors. The amount of candidate precipitation factors of May in area 1, area 2 and area 3 are 26, 19 and 30, respectively. In this paper, the prediction was taken from the data ranging from 2003 to May, 2008 in the area 1.

In the coupling, non-linearity and information redundancy existing among prediction factors will disturb the model's prediction strategy, and the prediction model does not lead to an ideal estimating result. In this paper, utilizing KPCA, dimension reduction is carried out for the 26 precipitation factors which are selected through cluster investigation. Then we select 8 main integrative factors as final precipitation prediction factors. Take the 8 main factors as the input variables for SVM net, and establish the daily precipitation prediction model, which continuously predicts the precipitation of area 1 from 2003 to May, 2008. Meanwhile, we select 6 main factors as the input variables for SVM net, and establish the monthly precipitation prediction model for the time period from 2001 to 2006.

4.2 Normalization Progressing of Data

The dimensions and orders of magnitude of extracted prediction factors are different form one another, so they are not suitable for PCA. Thus, normalization progressing is in demand for them. The normalization progressing on

every dimension of extracted prediction factors can be carried out by using the following method.

$$S' = \frac{S - S_{min}}{S_{max} - S_{min}} \tag{17}$$

In the above formula, S' is the prediction factor value after the normalization progressing. S is the prediction factor value before the normalization progressing. S_{min} is the minimum value of prediction factor values. S_{max} is the maximum value of the prediction factor values. By using this way, the relative maximum response value, absolute maximum response value, mean value and the curve's data fitting parameter of every prediction factor are all in the range of [0, 1], which is in favor of later data progressing.

4.3 SVM Parameter Optimization Based on PSO

Because the kernel parameter and error penalty parameter may have a great impact on the prediction performance of SVM, optimization for the two parameters is of great importance. The PSO is a global optimization methods based on swarm intelligence. It is excellent in global optimization and particularly suitable for the selection and optimization of model parameters. Therefore, in this paper the PSO is adopted to optimize SVM parameter and to figure out the optimal kernel parameter and the error penalty factor. We select the minimum mean square error (MSE) as fitness function.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i' - y_i)^2 \tag{18}$$

In the above formula, n is the number of prediction samples. y_i and y_i' are the measured value and the predicted value of No.i prediction sample, respectively.

During the progress of particle optimizing, every particle stands for a potential optimal solution for the extreme value optimization problem. The essential characteristics of particle are described by three indexes, including location, speed and fitness value. The speed decides the direction and distance of particle's movement. The fitness value is calculated through fitness function and it decides whether the particle is good or bad. In every circulation, the particle updates according to individual optimum and global optimum. The specific steps of PSO optimizing SVM parameter can be found in [13–15].

4.4 Establishment of Model

There are roughly four steps for the establishment of precipitation model. Step 1, use clustering analysis to handle the regional prediction and the prediction factors are extracted from fields general investigation. Step 2, extract the nonlinear characteristic factors from precipitation system by KPCA. Step 3, optimize the kernel parameter σ and error penalty factor c through PSO. Step 4, take the calculated optimization parameter value as SVM's optimal learning parameter to predict the samples. The establishment of model is shown in Fig. 1.

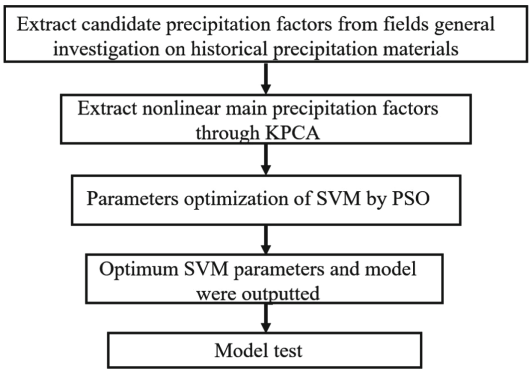

Fig. 1. Establishing progress of SVM precipitation model.

5 Analysis on Application Examples

5.1 Simulation of Precipitation in Guilin, Guangxi on a Daily Basis

Take the area 1 as an example. The data was collected through 148 days from 2003 to May of 2007 in Guilin, Guangxi are used as model training sample. The 31-day data of the May of 2008 works as testing sample. Figure 2 is a working sketch of the fitness of testing samples' data by improved RBF net model, KPCA-SVM model and KPCA-PSO-SVM model, respectively. As depicted in the figure, the testing result of KPCA-PSO-SVM model is generally in consistent with actual data, which indicates that among the three models, KPCA-PSO-SVM model has the best predicting performance, minimum total deviation and highest accuracy.

In order to analyze the predicting results more comprehensively, the following four evaluation indexes are introduced in this paper: mean absolute error (MAE1), maximum absolute error (MAE2), frequency of errors, which are greater than 25 mm (F1), and frequency of errors, which are smaller than 5 mm (F2). The error comparison between predicting results of the three models and T213 numerical prediction is listed in Table 1.

As demonstrated in Table 1, both MAE1 and MAE2 of KPCA-PSO-SVM model are smaller than those of the other four models, which proofs the high accuracy of KPCA-PSO-SVM model. Suppose that predictions with the errors being smaller than 5 mm are the reference value, and those with error being greater than 25 mm are unreliable prediction. Then we can find that the duration of the reference value using KPCA-PSO-SVM model and RBF net model is 17 days, which is longer than that of the other two models. However, the unreliable frequencies of KPCA-PSO-SVM model and RBF net model are 1 and 2, respectively, which means that KPCA-PSO-SVM model has better prediction performance.

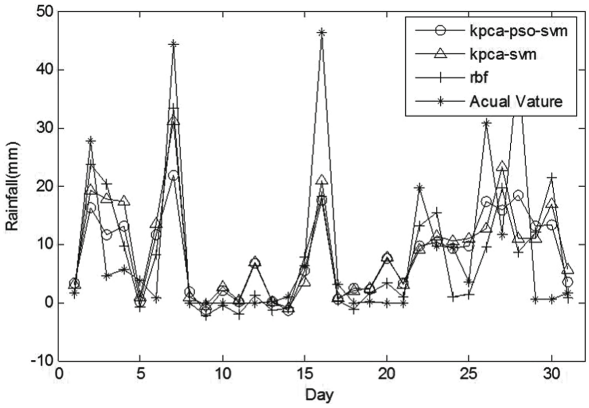

Fig. 2. Daily prediction rendering of area 1 during the 31 days in May.

Table 1. Analysis on comparison between the three prediction models and T213 numerical prediction.

Error statistics	RBF	KPCA-SVM	T213	KPCA-PSO-SVM
MAE1	32.7	29.8	36.1	28.7
MAE2	6.90	7055	7.92	6.81
F1	2	2	5	1
F2	17	13	15	17

5.2 Simulation of Precipitation in Guilin, Guangxi on a Monthly Basis

To verify the model's generalizing and stabilizing ability, the KPCA-PSO-SVM model proposed in this paper is applied in the monthly precipitation prediction of Guilin, Guangxi. It will be compared to the prediction of KPCA-SVM model and the improved RBF neural net model. Monthly precipitation data from 2001 to 2016 in Guilin, Guangxi is used in the simulation, including simulating data of the 156 months from 2001 to 2013 and testing data of the 36 months from 2014 to 2016.

Figure 3 lists the comparison between the monthly precipitation prediction results of the three models and the actual data, namely improved RBF net model, KPCA-SVM model and KPCA-PSO-SVM model. The simulating result shows that the testing simulation's data trends of these models are almost identical to the actual data tend. Among the three models, the KPCA-PSO-SVM model has the smallest deviation and better consistency, and it can be regarded an important reference for the protection against flood and drought.

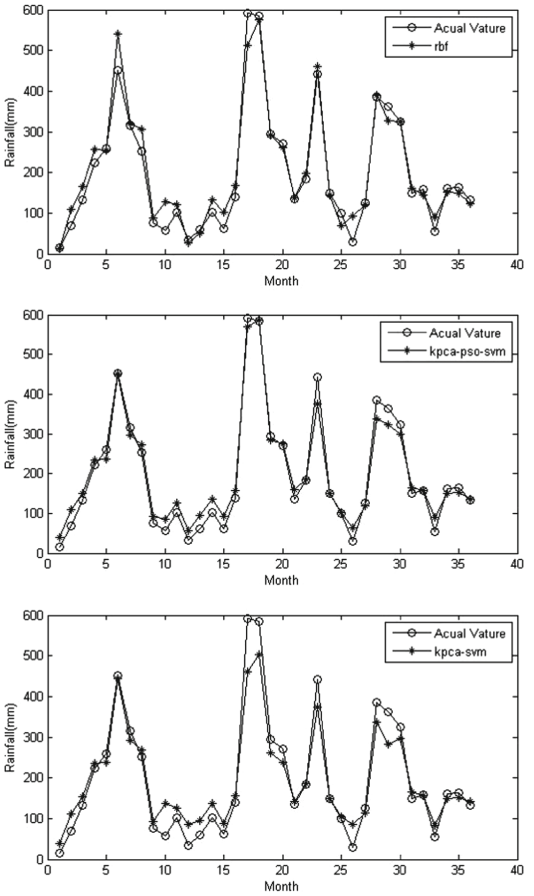

Fig. 3. The forecasting effect of three models between 2014–2016.

6 Conclusions

The accuracy of precipitation prediction is an important research topic for disaster reduction and prevention. With rapid development of economy and science technology, the requirement for high climatic prediction accuracy becomes increasingly higher. However, the combined influence of general atmospheric circulation and local circulation changes brings great difficulties in prediction. In this paper, based on the data from China Meteorological Administration and the Japanese refined net prediction, we adopted KPCA to extract the precipitation weather factors to optimize the SVM parameter through PSO to establish the daily precipitation prediction model in Guangxi. Meanwhile, it is applicable to the monthly precipitation prediction of Guilin, Guangxi. Simulation results show that in both aspects of maximum prediction error and mean prediction error, KPCA-PSO-SVM prediction achieves higher accuracy and shows better

generalizing ability than other methods. It demonstrates desirable stability and can work as a good reference for practical precipitation prediction.

References

1. Toda, Y., Abe, F.: Prediction of precipitation sequences within grains in 18Cr-8Ni austenitic steel by using system free energy method. ISIJ Int. **49**(3), 439–445 (2009)
2. Ouyang, Q., Wenxi, L., Dong, H., et al.: Study on precipitation prediction based on SVM regression analysis. Water-Saving Irrig. **9**, 38–41 (2014)
3. Ni, Y.: Study on Donggang precipitation prediction model based on SVM. Water Conservancy **2**, 133–134 (2014)
4. Teng, W., Shanxian, Y., Bo, H., et al.: Application research of SVM regression in flood and drought prediction in flood season. J. Zhejiang Univ. (Science) **35**(3), 343–347 (2008)
5. Yang, L., Xiwen, L., Liu, P., et al.: Time sequence analysis and the application of Monte Carlo in precipitation prediction. Environ. Sci. Technol. **34**(5), 108–112 (2011)
6. Sun, M., Kong, X., Geng, W., et al.: Time sequence analysis on shandong monthly precipitation based on ARIMA model. J. Ludong Univ. (Natural Science) **29**(3), 244–249 (2013)
7. Zhou, Z., Xie, B.: Application of BP neural net in Zhenzhou drought prediction and strategies of disaster reduction and prevention. Chin. Rural Water Conservancy Hydroelectricity **12**, 97 (2011)
8. Tao, W., Hui, Q., Li, P., et al.: Application of weighting Markov chain in precipitation prediction of Yinchuan area. South-to-North Water Divers. Water Sci. Technol. **8**(1), 78–81 (2010)
9. Liu, D., Fu, W.: Prediction test of least squares SVM in precipitation of flood season. In: The 33rd Annual Meeting of Chinese Meteorological Society S1 Supervision, Analysis and Prediction of Disaster Whether. Publishing House, Xi'an, pp. 929–934 (2016)
10. Gao, X.: Kernel feature extraction method and its application research. Nanjing University of Aeronautics and Astronautics (2010)
11. Luo, F., Jiansheng, W., Jin, L.: Integrated precipitation prediction model based on least squares SVM. J. Tropic. Meteorol. **27**(3), 577–584 (2011)
12. Luo, F.: Optimize RBF neural net precipitation prediction model based on LLE. Comput. Digit. Eng. **41**(5), 749–752 (2013)
13. He, X., Wang, Y., Wen, B.: Quantitative research on special engineering costs based on PSO SVM. Electricity Grid Clean Energy **31**(12), 27–30 (2015)
14. Li, T., Zeng, X.: Simulation of flow and sediment of yanhe basin based on PSO SVM. J. Basis Sci. Eng. **23**(7), 79–87 (2015)
15. Meng, J.: Study on long-term precipitation prediction model for arid region based on PSO-LSSVM. J. Yangtze River Sci. Res. Inst. **10**, 36–40 (2016)