第11卷 第23期 2011年8月
1671—1815(2011)23-5506-06

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 11 No. 23 Aug. 2011

地球科学

基于支持向量回归机的气象观测站优化模型

丁 兰 贾友见
(昆明理工大学理学院,昆明 650500)

摘 要 在保证足够信息量的前提下,针对合理减少气象观测站的实际问题,首先利用主成分分析(PCA) 降低样本数据的维数。其次利用支持向量回归机(SVR)对样本进行有效的回归。然后结合优化软件 lingo 对凸二次规划问题(与支持向量回归机相对应)进行求解。最终得出基于主成分分析-支持向量机回归预测优化模型。

关键词 回归 主成分分析法 支持向量回归机

中图法分类号 P413; **文献标志码** A

Optimization of meteorological observation station in real life has become a hotspot. Two methods following often used to forecast the precipitation: numerical prediction method is based on meteorology principle creating a series of partial differential equations, and getting the results of the equation with initial field to obtain the predicted results; probabilistic method is statistical law of analyzing the evolution of weather and numerical relationship of analyzing prediction factor and predict. There is established mathematical model to predict future weather. Neural network have strong ability to deal with nonlinear problems and the method is also widely used to predict field. In practice, there being found the methods above is very well to fit known date, but bias are often found when predict unknown samples and its more serious in small sample. This phenomenon is known as poor generalization ability in mathematical. In response to these problems, Vapnik have created a machine learning algorithm based on statistical learning-support vector machine[1—3] in 90s' of mid – 20th century.

Parameter selection is one of the research focuses[4—6] of Support Vector Machine. Support Vector regressing samples efficiently are used and obtained parameter values. In order to improve the promotion ability of decision function to obtain the optimal value in support vector regression, can converse machine learning problem to convex quadratic programming problem. The simple Tools LINGO (Linear Interactive and General Optimizer) software which is often applied linear and nonlinear optimization problems are used getting the regression prediction model.

In some of complex prediction systems, the selected factors are likely serious correlation. This correlation sometimes affects the effect of prediction[7] seriously. PCA can effectively solve the multi-correlation problem among variables, using SVM training samples and LINGO solving it, and obtaining optimal regression model.

1 SVR prediction model based on principal

1.1 Principal component analysis

1.1.1 Principal component analysis process

Principal component analysis is a comprehensive

2011年4月28日收到,5月10日修改 国家自然科学基金(10847139)、云南省科学基金(2009CD036、08Z0015)资助
第一作者简介:丁 兰(1988—),湖北人,研究生。研究方向:应用数学。E-mail: 184776258@163.com。

multivariate statistical analysis method using dimension reduction to transfer original multi-index a few independent composite index, which are widely used in forecasting and analysing atmospheric science research.

Definition[2] contribution rate of first principal component is known as $t_i = \lambda_1 / \sum_{i=1}^{p} \lambda_i$, as a result of $\mathrm{Var}(F_1) = \lambda_1, \lambda_1 / \sum_{i=1}^{p} \lambda_i = \dfrac{\mathrm{Var}(F_i)}{\sum_{i=1}^{p} \mathrm{Var}(F_i)}$. Therefore, the contribution rate of first principal component is ratio of first principal component of the variance divided by total variance $\sum_{i=1}^{p} \lambda_i$. The larger the ratio indicate that the stronger the ability of integrated information $x_1, x_2, \cdots, x_p$ of the first principal component.

The cumulative contribution rate of the first k principal components is defined as $T_i = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$. If λ_i is arranged by small to large, and the cumulative contribution rate of the first k principal components is only 15% or less, then the remaining p-k principal components are included all the measuring indicators of information, there are not only reduces the number of variables, but also facilitate the analysis and study of practical problems.

There are 10 observation station in one city, want to reduce the number of weather stations to save money, and what stations can not only reduce the cost savings but also ensures that the information of annual precipitation of the city is large enough (*i. e.* information loss is little as possible). Purpose of this article is to obtain the annual precipitation of the reduced weather stations, and the date is observed by others stations after reduced some of the stations. Table 1 has the annual precipitation data of 10 stations in the past 30 years, which have $n = 10$ samples and $p = 30$ sample values in each observed sample.

Table1　1976—2005 annual precipitation measured in each observation station (Unit/mm)

date	meteorological observation station									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1976	600	464	584	448	648	176	328	232	488	544
1977	488	384	520	416	432	432	536	448	512	448
1978	616	520	616	488	544	504	536	496	432	592
1979	688	440	520	352	880	376	456	432	552	440
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2002	824	552	672	472	472	424	408	360	512	576
2003	688	584	680	584	432	416	392	376	568	440
2004	744	416	480	472	504	432	328	320	576	368
2005	624	520	680	424	672	536	632	544	336	584

Original data matrix is obtained by the principal component analysis:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_p \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (X_1, X_2, \cdots, X_P) \tag{1.1.1}$$

Where $X_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix}, i = 1, 2, \cdots, p.$

1.1.2　Principal components determination

Descript original data to matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1.1.2}$$

Calculated the correlation coefficient matrix above

$$R = (r_{ij})_{p \times p} \tag{1.1.3}$$

Solving the eigenvalue $0 \leqslant \lambda_1 \leqslant \lambda_2 \leqslant \cdots \lambda_p$ of R and corresponding eigenvalue vector, contribution rate, the cumulative contribution rate.

Characteristic values $\lambda_1 = 0.032\ 1$, $\lambda_2 = 0.057\ 1$, $\lambda_3 = 0.093\ 3$, can be seen from table 2 are approximately equaes 0, only the first eight components of the absolute value in v_1 is the maximum value in the first calculated results and the corresponding eigenvector v_1, v_2, v_3. Without loss of information can delete station X_8 in this case. Because of the correlations of the stations, one station can only be deleted each time. Then delete X_3 based on the method above with the date of the remaining 9 stations in 30 years. Finally, delete X_6 with the remaining 8 stations following the method above. In summary, characteristics of stations X_3, X_6, X_8 are extracted based on the method of principal component analysis. To analyse the precipitation of stations (table 2), obtained that correlation coefficient of X_7 and X_8 is 0.952 268, with a high degree of correlation. In which, X_2 and X_3, X_6 and X_8, X_7 and X_8 have strong correlation respectively. These illustrated that multicollinearity is exist in there; therefore, the article considered using X_2, X_8 and X_8 to predict the X_3, X_6 and X_7.

Table 2 Characteristic value and contribution rate of annual precipitation in each observation station

station	eigenvectors									
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}
X_1	0.051 3	0.010 8	0.016 8	−0.570 1	0.387 6	−0.169 2	−0.237 6	0.654 8	−0.003 3	0.100 6
X_2	−0.132 4	0.617 7	0.273	0.266 2	0.013 2	0.044 9	0.254 2	0.270 1	0.475 5	0.299 4
X_3	0.108 4	−0.670 4	−0.194 5	0.231 8	0.169 8	0.037 9	0.134 8	0.131 9	0.545 9	0.288 6
X_4	0.035 3	0.145 9	−0.116 3	0.489 4	0.252 8	−0.070 6	−0.801 9	−0.048 4	0.005 7	0.100 6
X_5	0.013 2	0.005 1	0.019 1	0.342 5	0.707 6	0.013 6	0.419 7	0.037 4	−0.449 9	−0.035 4
X_6	−0.290 4	−0.316 6	0.674 7	0.018 7	−0.099 1	0.112	−0.137 7	−0.027 2	−0.290 4	0.483 2
X_7	−0.506 1	0.106 1	−0.630 2	−0.079 1	−0.090 6	0.013	0.076 2	−0.038 5	−0.215 6	0.516 8
X_8	0.785 8	0.130 9	−0.07	−0.025 8	−0.166 6	−0.070 6	0.070 4	−0.015 6	−0.238 3	0.514 9
X_9	−0.073 2	−0.135 3	−0.030 6	0.401 4	−0.414 6	−0.543 3	0.087 2	0.512 9	−0.232 2	−0.150 4
X_{10}	−0.076 3	0.025 4	0.123 4	−0.158 7	0.196	−0.806 2	0.064 8	−0.46	0.178 3	0.131 4
eigenvalue λ_i	0.032 1	0.057 1	0.093 3	0.560 1	0.592	0.828 6	1.222 5	1.409 2	2.057 1	3.148 1
contribution rate t_i	0.003 2	0.005 7	0.009 3	0.056	0.059 2	0.082 9	0.122 3	0.140 9	0.205 7	0.314 8
cumulative contribution rate T_i	0.003 2	0.008 9	0.018 2	0.074 2	0.133 4	0.216 3	0.338 6	0.479 5	0.685 2	1

1.2 SVR prediction model

1.2.1 SVR optimization problem

Training set: $T = \{(x_1, y_1), \cdots (x_l, y_l)\} \in (R^n \times y)^l$, where, $x_i \in x = R^n$ is input index, $y_i \in y = R, i = 1, \cdots, l$ is output indicators. Tried based on it to find a real-valued function $g(x)$ in R^n and conveniently to infer the corresponding output y of any input value x, the goal is to search a smooth curve $y = g(x) = (w \cdot x) + b$, which close to the input points.

Original optimization problem of linear regression function $y = (w \cdot x + b)$:

$$\min \frac{1}{2} w^2 \tag{1.2.1}$$

$$\text{s. t.} \quad (w \cdot x_i) + b - y_i \leqslant \varepsilon, i = 1, \cdots, l \tag{1.2.2}$$

$$y_i - (w \cdot x_i) - b \leqslant \varepsilon, i = 1, \cdots l \tag{1.2.3}$$

For the regression of one-dimensional R of regression line, constraining equaes (6) and (7) means that all training points " × " should be within the regression line ε. Objective function means that the regression line should be the minimum slope line satisfying the

conditions above (figure 1).

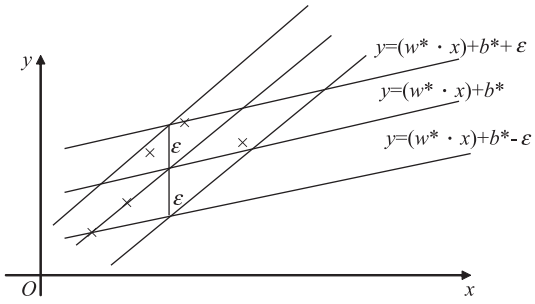

Fig. 1　Soft ε-band hyperplane

Introduced Slack variable[8] $\xi^{(*)} = (\xi_1, \xi_2 \cdots, \xi_l, \xi_l^*)$, penalty parameterand C and Lagrange function

$$L(w,b,\xi^{(*)},\alpha^{(*)},\eta^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i + y_i - (w \cdot x_i) - b) - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* - y_i + (w \cdot x_i) + b)。$$

Where, $\alpha^{(*)} = (\alpha_1, \alpha_1^*, \cdots, \alpha_l, \alpha_l^*)^{\mathrm{T}}$, $\eta^{(*)} = (\eta_1, \eta_1^*, \cdots, \eta_l, \eta_l^*)^{\mathrm{T}}$ is Lagrange Multiplier vector.

Then the convex quadratic programming problem of the linear ε support vector machine (1.2.1)—(1.2.3) is obtained:

$$\min_{\alpha^{(*)} \in R^{2l}} \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i x_j) + \varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i) \tag{1.2.4}$$

$$\text{s.t} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \tag{1.2.5}$$

$$0 \leqslant \alpha_i^* \leqslant C,\ i = 1, \cdots, l \tag{1.2.6}$$

Have solutions $\bar{\alpha}^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \cdots, \bar{\alpha}_l, \bar{\alpha}_l^*)^{\mathrm{T}}$.

1.2.2　Constructed decision function

Calculated $\bar{b}$: Selected the component $\bar{\alpha}_j^*$ or $\bar{\alpha}_k^*$ from $\bar{\alpha}_i^*$ locating the open interval$(0, C)$,

$$\bar{b} = y_j - \sum_{i=1}^{l}(\bar{\alpha}_i^* - \bar{\alpha}_i)(x_i x_j) + \varepsilon \tag{1.2.7}$$

If elected $\bar{\alpha}_k^*$ then:

$$\bar{b} = y_k - \sum_{i=1}^{l}(\bar{\alpha}_i^* - \bar{\alpha}_i)(x_i x_k) - \varepsilon \tag{1.2.8}$$

Decision function:

$$y = g(x) = \sum_{i=1}^{l}(\bar{\alpha}_i^* - \bar{\alpha}_i)(x_i x) + \bar{b} \tag{1.2.9}$$

1.2.3　SVR prediction

It can be known that some stations have multicollinearity each other by the principal component analysis. Used the observation stations X_2, X_8, X_8 to predict X_3, X_6, X_7, *i. e.* training set:

$x = \{(x_2, y_2), (x_8, y_8), (x_8, y_8)\}$, where, $x_i \in x = R^3$ are input index vectors, $y_i \in y = R, i = 1, \cdots, l.$ are output indicators. Using Matlab 2010 we get SVM parameters optimization cross-validation, minimum mean deviation $MSE = 0.315\ 83$. Through choosing[9] parameters by regression prediction model, obtained Loss function parameters $\varepsilon = 0.062\ 5$ ($g = \varepsilon$) and Penalty parameter $c = 0.574\ 35$, shown in fig. 2.

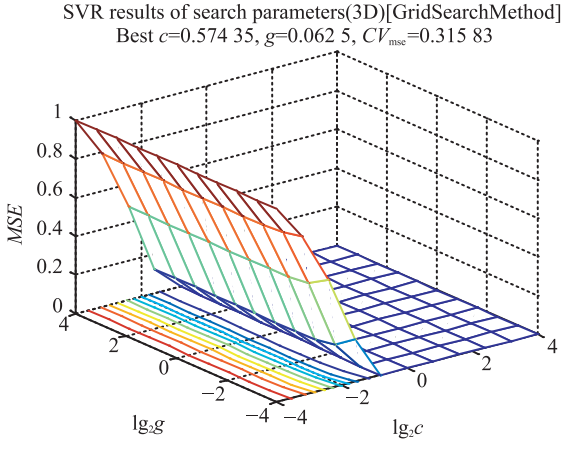

Fig. 2　Results of search parameters

1.2.4　Model solution

Convex quadratic programming duality problem (1.2.4)—(1.2.6), can be solved by optimization

software Lingo80, and received the results[10].

Local optimal solution found at iteration: 10

Objective value: -0.437 500 0

Variable	Value	Reduced Cost
A(1)	0.000 000	0.999 992 8
A(2)	0.000 000	0.605 988 0E-01
A(3)	0.500 000 0	0.000 000
B(1)	0.500 000 0	0.000 000
B(2)	0.000 000	0.644 048 0E-01
B(3)	0.000 000	0.125 010 8
Y(1)	1.000 000	0.000 000
Y(2)	0.606 060 0E-01	0.000 000
Y(3)	0.000 000	0.000 000

i.e. $\alpha = (0, 0.5, 0, 0, 0.5, 0)$, Substituting (1.2.7)—(1.2.9),

Get $\bar{b} = (0, 0.4375, 0, 0, 0.625, 0)$,

Decision function: $y = \frac{1}{2}(x_2 - x_8)x + b$.

Therefore, the comparison results between the original data and the regressive prediction data are shown in figure 3. SVM network regressive prediction results are received by SVM network training: correlation coefficient of the mean square error $MSE = 0.018\ 042\ 5$ is $R = 99.707\ 9\%$, the regressive prediction effect is well.

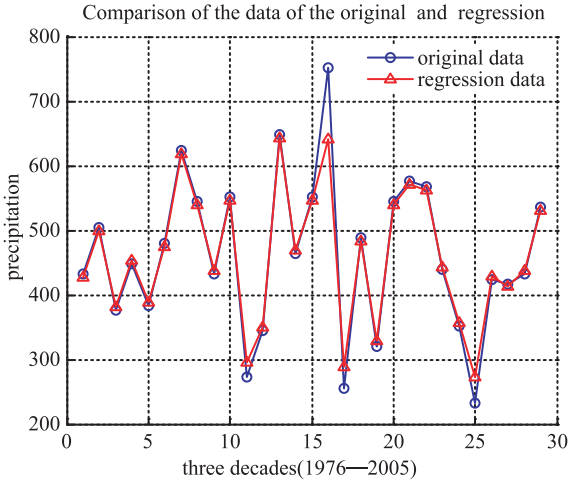

Fig. 3 Results of regression prediction

2 Conclusion

Regressive prediction model is established based on principal component analysis-support vector machine to solve the problem of reducing meteorological observation stations which was encountered in the field of meteorological observations. The model is a regression technique-support vector regression constructing based on support vector machine. Though parameters choosing and giving the method of decision-making model solution, the field of support vector machine theory and application are extended, and three actual samples are predicted with perfectly effect. Generalization ability and prediction accuracy of the SVR technique are confirmed excellently.

References

1 Deng N Y, Tian Y J. A new data mmining method-support vector machines. Beijing: National Defence Industry Press, 1989

2 Yu X, Yun X S. Multivariate statistical analysis. Beijing: China Statistics Press, 1999

3 Ancona N. Classification properties of support vector machines for regression. Technical Report, RI-IESI/CNR-Nr, 1999

4 Liu H C, Ma S Y. Research support vector machine. Journal of Image Graphics Newspaper, 2007; 7(6): 618—623

5 Feng Z H, Yang J M. Parameters of SVM regression. Mechanical Engineering and Automation, 2007; 3: 17—22

6 Guyon I, Weston J, Barnhill S, *et al*. Gene selection for cancer classification using support vector machines. Machine Learning, 2000; 46(13): 389—422

7 Sun Z, Yang P. Gene expression profiling on lung cancer outcome prediction: present clinical value and future premise. Cancer Epidemiology Biomarkers &Prevention, 2006; 15(11): 2063—2068

8 Castro-Neto M, Jeong Youngseon, Jeong M K, *et al*. HanAADT prediction using support vector regression with data-dependent parameters. Expert Systems with Applications, 2009; 36(2): 2979—2986

9 Rajasekaran S, Gayathri S, Lee T-L. Support vector regression methodology for storm surge predictions. Ocean Engineering, 2008; 35(16): 1578—1587

10 Xue Y, Geng M Y. Operations research and experimental. Beijing: Electronic Industry Press, 2008

Optimization Model of Meteorological Observation Station Based on Support Vector Regression

DING Lan, JIA You-jian

(School of Science, Kunming University of Science and Technology, Kunming 650500, P. R. China)

[**Abstract**] In the premise to ensure enough information, some measures are taken to reduce meteorological observation station properly. Dimensions of samples are reduced with the principal component analysis method, then samples are regressed effectively based on support vector regression (SVR) finally the regressive forecast model is established combining with Linear Interactive and General Optimizer to solve the convex quadratic programming corresponding to support vector regression.

[**Key words**] regression　principal component analysis　support vector regression

(上接第 5505 页)

参　考　文　献

1　罗振东. 有限元混合法理论基础与应用. 济南,山东教育出版社,1996

2　王同科. 几类微分方程数值算法研究,山东大学博士学位论文. 济南,山东大学数学与系统科学学院,2002

3　李　宏,刘　洋. 一类四阶抛物型积分-微分方程的混合间断时空有限元法. 计算数学,2007;29(4):414—420

4　纪维强,杨　青. 四阶半线性抛物型方程(组)的混合体积元方法及其数值模拟,山东师范大学硕士学位论文. 济南,山东师范大学数学科学学院,2009

Mixed Finite Volume Element Method for the Four Order Parabolic Integro-differential Equations

CONG Mei-qin, YANG Qing

(Department of Mathematics Science, Shandong Normal University, Jinan 250014, P. R. China)

[**Abstract**] The mixed finite volume element method is used for the four order parabolic integro-differential equation with the initial-value problems, and the error estimates of semi-discrete solutions are obtained.

[**Key words**] mixed finite volume element　fourth order parabolic integro-differential problems　error estimate