KPCA_SVM 水文时间序列预测模型的建立与应用

邵年华[1],沈 冰[1],黄领梅[1],戴玉萍[2]

(1 西安理工大学 西北水资源与环境生态教育部重点实验室,陕西 西安 710048;2 新疆和田河管理局,新疆 和田 848000)

[摘 要] 【目的】建立水文时间序列预测的核主成分支持向量机(KPCA_SVM)模型。【方法】利用核主成分分析(KPCA)对输入数据进行非线性特征信息提取,并将提取的特征信息作为最小二乘支持向量机(LSSVM)的输入变量,建立 KPCA_SVM 预测模型。以甘肃民勤地区的月蒸发量为例,对模型的预测效果进行检验。【结果】预测结果表明,KPCA_SVM 模型预测效果优于 PCA_SVM 模型和 LSSVM 模型,预测平均相对误差为 8.36%。【结论】KPCA_SVM 模型的预测效果优于没有特征提取的 LSSVM 模型。与主成分分析(PCA)提取特征相比,KPCA 特征提取效果更好。

[关键词] 水文时间序列;蒸发量;核主成分分析;支持向量机;KPCA_SVM 模型

[中图分类号] P333.1 [文献标识码] A [文章编号] 1671-9387(2009)09-0204-05

Establishment and application of hydrological time series forecasting model based on KPCA_SVM

SHAO Nian-hua[1], SHEN Bing[1], HUANG Ling-mei[1], DAI Yu-ping[2]

(1 *Xi'an University of Technology, Key Lab of Northwest Water Resources and Environmental Ecology, MOE, Xi'an, Shaanxi* 710048, *China*; 2 *Administration Bureau of the Hotan River, Hotan, Xinjiang* 848000, *China*)

Abstract: 【Objective】 The KPCA_SVM model of hydrological time series forecasting model was established. 【Method】 The method of Kernel Principle Component Analysis (KPCA) was used to obtain the feature information, and then the obtained series was used as input of Least Square Support Vector Machine model for forecasting. With monthly evaporation in the Minqin region as an example, it was applied to test forecasting result of model. 【Result】 The results show that the KPCA_SVM model had a better effect on forecasting than PCA_SVM and LSSVM, and the average error was 8.36%. 【Conclusion】 The forecasting effect of KPCA_SVM model was much better than that of LSSVM model without obtaining the feature information. In comparison with PCA, the performance of KPCA was better, too.

Key words: hydrological time series; evaporation; kernel principle component analysis; support vector machine; KPCA_SVM model

水文系统是一个复杂的非线性系统,其所含的要素之间也存在着非常复杂的非线性关系。因此,要想充分揭示水文系统的特性,准确地把握系统中要素的变化规律极为重要。目前,对水文要素(如降雨、蒸发、径流等)的预测有两大途径:一是基于要素自身变化规律建立预测模型;二是基于要素及其影响因素之间的关系建立预测模型[1]。虽然利用第 1 种途径对水文要素进行预测已经取得了一定成果,但在实际的水文系统中,一个要素的变化往往受到其他多个因素的影响,所以要想更好地揭示水文系

* [收稿日期] 2008-12-20
[基金项目] 国家自然科学基金项目(50779052)
[作者简介] 邵年华(1983-),男,辽宁大连人,在读硕士,主要从事旱区水文水资源研究。E-mail:snhsnhpop@yahoo.com.cn
[通信作者] 沈 冰(1948-),男,浙江湖州人,教授,博士生导师,主要从事旱区水文过程及水资源演变研究。E-mail:shenbing@xaut.edu.cn

统的特性，仅仅通过单要素进行建模预测是不够的。许多学者也充分认识到这个问题，因此近年来基于多变量的水文要素预测方法以及基于相空间重构，将一维时间序列变换成多维序列进行水文要素预测的方法得到了广泛应用，并取得了一定的成果。但在实际应用中，并不是考虑的因素越多预测效果就越好，因素过多，由于变量之间往往存在着复杂的相关性，因此很难直接抓住其间的主要关系，这就需要对数据进行简化，使高维数据降维，从而获得数据的主要信息，提高预测模型识别的准确率，减少识别的工作量。主成分分析（Principle component analysis，PCA）为输入数据降维提供了一种很好的方法，但 PCA 方法的核心过程是采用一组线性变换进行空间映射[2]，而水文系统中各要素之间在本质上是非线性关系，因此 PCA 不适用于水文时间序列的降维。核主成分分析（Kernel principle component analysis，KPCA）作为一种非线性的 PCA 方法，可以有效地提取输入数据的非线性信息[3]，因而在非线性的特征提取、分类和模式识别等方面得到了广泛应用[4-6]，但该方法至今在水文领域应用较少。支持向量机（Support vector machine，SVM）是由 Vapnik 在 1995 年提出的新型统计学习方法[7]，模型结构简单，对小样本数据有很好的泛化能力，目前已在很多领域得到了成功应用[8-10]。SVM 在构造最优决策函数时，运用结构风险最小化原则，并利用核函数取代内积运算，较好地解决了非线性、高维数等问题[11]。但 SVM 也存在一些不足，在处理大样本或相关性复杂的数据时，计算量增加，导致训练速度较慢，精度较差。

本研究将 KPCA 方法与 SVM 理论结合，建立 KPCA_SVM 模型，利用 KPCA 提取输入数据特征，获得数据间的主要信息，去除重复相关性，这样既可以弥补 SVM 在训练时的不足，又可以发挥两种方法各自的优点。本研究同时还探讨了 KPCA_SVM 模型在甘肃民勤地区蒸发量预测中的应用，通过与其他方法的对比，分析该模型的应用前景，以期为水文要素预测提供理论依据。

1　核主成分分析[12]

基于核函数的 PCA 方法不是直接计算特征向量，而将是将其转化为求核矩阵的特征向量和特征值，避免了在特征空间求特征向量，而数据在特征向量上的投影转换为求核函数的线性组合，使计算大大简化。

首先给定一个样本 x_k：$k=1, 2, \cdots, n$，$x_k \in R^N$，n 为样本总数。将其映射到特征空间 $\Phi(x_k)$：$k=1, 2, \cdots, n$，$x_k \in R^N$，$\Phi(\cdot)$ 是一个非线性映射。计算协方差矩阵 C：

$$C=\frac{1}{n}\sum_{j=1}^{n}\Phi(x_j)\Phi(x_j)^T, j=1,2,\cdots,n \text{ 。} \quad (1)$$

然后通过解特征值问题计算主成分，可以找到当特征值 $\lambda>0$ 和特征向量 $V\neq 0$ 时，满足：

$$\lambda V=CV。 \quad (2)$$

对应非零特征值的特征向量 V 可由映射到特征空间的样本矢量线性表示为：

$$V=\sum_{j=1}^{n}\alpha_i\Phi(x_i)。 \quad (3)$$

式中：α_i 为方程系数，$i=1,2,\cdots,n$。

式(2)左乘 $\Phi(x_k)$ 变为：

$$\lambda(\Phi(x_k),V)=(\Phi(x_k),CV), k=1,2,\cdots,n。 \quad (4)$$

定义一个 $n\times n$ 矩阵 K_{ij}，即：

$$K_{ij}=K(x_i,x_j)=(\Phi(x_i),\Phi(x_j)), \quad i,j=1,2,\cdots,n。 \quad (5)$$

计算系数 α_i 可转变为求核矩阵 K 的特征向量和特征值，即：

$$n\lambda\alpha=K\alpha。 \quad (6)$$

式中：α 为 $\alpha_i(i=1,2,\cdots,n)$ 所组成的列向量。归一化特征向量 V，此时样本 $\Phi(x_k)$ 在 V 上的映射为：

$$h_i(x)=(V,\Phi(x))=\sum_{i=1}^{n}\alpha_i^i\Phi(x_i)\Phi(x)= \sum_{i=1}^{n}\alpha_i^iK(x_i,x), i=1,2,\cdots,n。 \quad (7)$$

式中：$h_i(x)$ 为对应于 $\Phi(x)$ 的第 k 个非线性主成分分量。

特征值 λ_i 小的主成分 h_i 可以认为是噪声引起的，比值 $\lambda_i/\sum_{i=1}^{n}\lambda_i$ 反映了分量 h_i 对整体方差的贡献，较重要的分量对应较大的比值。主成分数量的选取原则为：

$$[\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{n}\lambda_i]>E。 \quad (8)$$

式中：m 为选取的主成分数量；E 为选取的百分比值，通常 $E>85\%$。

以上推导的特征空间变量均值均是以 $\sum_{i=1}^{n}\Phi(x_i)=0$ 为假设条件的，然而实际中的样本数据并不一定满足 $\sum_{i=1}^{n}\Phi(x_i)=0$，此时式(6)中 K 的取值为：

$$\overline{K}=K=lK=Kl+lKl。 \quad (9)$$

式中：l 为系数，是 $1/n$ 的 $n\times n$ 阶单位矩阵。

2 最小二乘支持向量机

训练数据为 n 个样本,可以表示为:$x_k:k=1,2,\cdots,n,x_k\in R^N$;$y_k:k=1,2,\cdots,n,y_k\in R^L$。在特征空间中,支持向量机的最优决策函数为:

$$y_k=w^T\Phi(x)+b。\tag{10}$$

式中:$\Phi(\cdot)$是输入空间 R^N 到高维特征空间 F 的非线性映射;w 为权矢量,$w\in F$;b 为函数偏置量。w,b 根据结构风险最小化原则,通过下式来估计。

$$Q=\frac{1}{2}\|w\|^2+c\zeta。\tag{11}$$

式中:Q 为正则化风险;$\|w\|^2$ 为控制模型的复杂度,c 为正则化参数,ζ 为允许错分的松弛变量。

用作函数逼近的最小二乘支持向量机的优化问题转化为:

$$\min J(w,\zeta)=\frac{1}{2}w^Tw+c\sum_{k=1}^{n}\zeta_k^2。\tag{12}$$

约束条件:$y_k=w^T\varphi(x_k)+b+\zeta_k,k=1,2,\cdots,n$。

最小二乘支持向量机优化问题对应的拉格朗日函数 L 为:

$$L(w,b,\zeta,\alpha)=J(w,\zeta)-\sum_{k=1}^{n}\alpha_k\{w^T\varphi(x_k)+b+\zeta_k-y_k\}。\tag{13}$$

式中:$\alpha_k(k=1,2,\cdots,n)$为拉格朗日乘子。利用拉格朗日函数分别对 w,b,ζ,α_k 求导,可得:

$$\begin{cases}\dfrac{\partial L}{\partial w}=0\Rightarrow w=\sum\limits_{k=1}^{n}\alpha_k\varphi(x_k),\\ \dfrac{\partial L}{\partial \zeta}=0\Rightarrow \alpha_k=c\zeta_k,\\ \dfrac{\partial L}{\partial b}=0\Rightarrow \sum\limits_{k=1}^{n}\alpha_k=0,\\ \dfrac{\partial L}{\partial \alpha_k}=0\Rightarrow w^T\varphi(x_k)+b+\zeta_k-y_k=0。\end{cases}\tag{14}$$

定义核函数 $k(x_k,x_i)=\varphi(x_k)\varphi(x_i)$,并根据(14)式将优化问题转化为求解如下线性方程问题:

$$\begin{pmatrix}0 & 1 & 1\\ 1 & k(x_1,x_1)+1/c & k(x_1,x_i)\\ \vdots & \vdots & \vdots\\ 1 & k(x_k,x_1) & k(x_n,x_n)+1/c\end{pmatrix}\begin{pmatrix}b\\ a_1\\ \vdots\\ a_n\end{pmatrix}=\begin{pmatrix}0\\ y_1\\ \vdots\\ y_n\end{pmatrix}。\tag{15}$$

最后得到方程:

$$f(x)=\sum_{k=1}^{n}\alpha_k k(x,x_k)+b。\tag{16}$$

核函数常取径向基函数[13]:

$$k(x_k,x_i)=\exp\left(\frac{-\|x_k-x_i\|_2^2}{\sigma^2}\right)。\tag{17}$$

式中:σ 为径向基函数的宽度,为待定参数。

3 KPCA_SVM 的建模过程

先利用 KPCA 对数据进行预处理,消除变量之间的相关性,提取主成分,降低数据维数;再利用最小二乘支持向量机(LSSVM)对提取的主成分进行训练,并调整参数;最后用检验集对模型进行检验,选取最佳参数组合作为模型的最终参数,具体建模步骤如下[14]。

(1)选取模型数据输入输出样本集 $x_k:k=1,2,\cdots,n,x_k\in R^N$;$y_k:k=1,2,\cdots,n,y_k\in R^L$(一般 L 取 1,即单变量输出),对样本数据进行归一化处理。

(2)利用公式(1)~(7)对输入样本数据进行非线性主成分分析,再进行核主成分提取,按照主成分贡献率由大到小选取,根据公式(8)选取累积贡献率达到 90%左右的 m 个主成分,其余可认为是噪声成分。

(3)将选取的 m 个主成分作为最小二乘支持向量机的输入样本,选取前 l 组数据作为训练集,剩余数据作为检验集。

(4)选取径向基函数作为最小二乘支持向量机的核函数,设定正则化参数 c 和核参数 σ 的范围,利用交叉验证法确定最优参数,并利用所选参数进行最小二乘支持向量机的训练。

(5)将检验集输入最小二乘支持向量机模型中进行检验,若误差不满足要求则返回上一步重新选取最佳参数组合,直到精度满足要求为止。

4 KPCA_SVM 模型的应用

甘肃民勤地区是典型的绿洲灌溉农业区,近年来随着人口的持续增加和工农业生产的不断发展,该区的生态环境,特别是水环境遭到严重破坏,主要表现为地下水位下降,土地荒漠化、盐渍化加重,植被退化,生态环境变得极其脆弱[15]。因此,对民勤地区蒸发量变化进行研究,可以深入了解该地区的气候变化规律,对该区的农业生产、生态环境治理和经济发展均具有十分重要的意义。本研究以民勤地区的月蒸发量为例,对 KPCA_SVM 模型的预测效果进行验证。

4.1 主成分的提取及主成分个数对模型精度的影响

本研究所用资料来自石羊河流域管理局。选取民勤地区 1991~2001 年共 132 个月实测的平均气温、相对湿度、降水、日照时数、风速等气象资料,作

为模型训练与检验的输入数据。取 1991～1998 年共 96 个月的气象资料作为训练集,依次编号为 1～96;1999～2001 年共 36 个月的气象资料作为检验集,依次编号为 97～132。用 KPCA 法对模型输入数据进行核主成分分析,KPCA 的核函数选用高斯核函数,按累积方差百分比大于 90%选取 4 个主成分作为 LSSVM 的输入量。为了更直观地了解主成分数对模型精度的影响,本研究给出了 KPCA 提取的主成分数与模型检验结果均方根误差(RMSE)之间的关系(表 1)。

表 1　KPCA 提取的主成分数与模型检验结果 RMSE 的关系

Table 1　Relationship between number of extracted principal components and testing results RMSE of model

主成分数 Principal component number	均方根误差 RMSE	主成分数 Principal component number	均方根误差 RMSE
1	4.044 4	4	0.695 9
2	4.118 9	5	0.946 4
3	2.866 7		

由表 1 可见,随着 KPCA 提取主成分数的增加,模型的 RMSE 呈下降趋势,主成分数为 4 时达最小,当主成分数增加到 5 时,RMSE 又有小幅增大。这说明 KPCA 提取主成分可以消除输入数据中的噪声,降低数据维数,从而提高预测模型的预测精度,最佳数量的主成分可以很好地代替原始数据;也说明主成分的个数并不是越多越好,多余的主成分反而会夹带更多的噪声,影响模型的预测精度。

4.2　KPCA_SVM 蒸发模型的建立及结果分析

将 KPCA 提取的 4 个主成分作为 LSSVM 的输入量进行训练,利用交叉验证选定正则化参数 c 为 30,核参数 σ 为 7.7。将检验集输入 LSSVM,得到的月蒸发量预测结果如图 1 所示。从图 1 可以看出,KPCA_SVM 的预测值与实测值拟合效果较好,说明利用所建立的 KPCA_SVM 蒸发模型可以很好地预测月蒸发量的变化趋势。误差分析结果表明,利用 KPCA_SVM 模型预测时的平均相对误差为 8.36%,最大误差为 20.7%,按水文规范要求以相对误差 20%为合格,则合格率为 97%(表 2)。

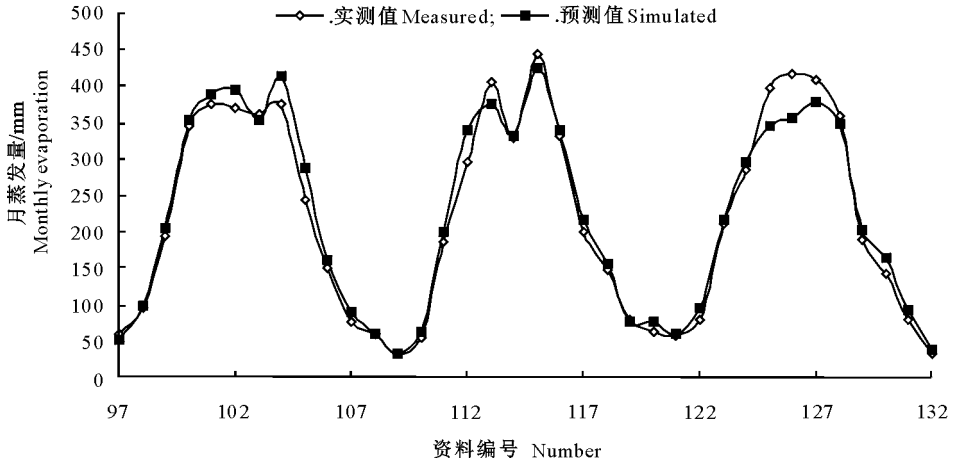

图 1　甘肃民勤地区月蒸发量的 KPCA_SVM 模型预测值与实测值

Fig. 1　Simulated and measured value of KPCA_SVM model for monthly evaporation in Minqin region, Gansu province

表 2　甘肃民勤地区月蒸发量 3 种模型的参数及其预测效果比较

Table 2　Parameters and the comparison of forecasting results for three models of monthly evaporation in Minqin region, Gansu province

模型 Model	参数 Parameter		平均相对误差/% Average relative error	最大相对误差/% Maximum relative error	误差>20%的个数 Number of error >20%	合格率/% Qualified rate
	c	σ				
LSSVM	65	49	10.2	48.4	4	89
PCA_SVM	60	2.5	15.7	59.2	10	72
KPCA_SVM	30	7.7	8.36	20.7	1	97

为了充分体现 KPCA_SVM 模型的学习和泛化能力,本研究选用 LSSVM 和 PCA_SVM 模型对相同资料进行预测,结果见表 2。由表 2 可看出,运用 KPCA 法对输入数据进行主成分提取后,再结合

LSSVM进行预测,效果要优于直接运用LSSVM进行预测,这充分说明原始数据中所夹带的噪音对模型预测精度影响很大。而运用PCA对输入数据进行主成分提取后的预测效果比LSSVM差,原因是干旱地区的月蒸发量变化较大,与影响因素之间的关系较为复杂,呈现出非线性关系,因此通过PCA对输入数据进行线性变化来达到除噪降维,往往不能获得令人满意的预测效果。

5 结 论

本研究结合KPCA和SVM,建立了水文时间序列的KPCA_SVM预测模型,并以甘肃民勤地区的月蒸发量为例,对模型预测效果进行验证,结果表明,该模型的预测效果较好;KPCA具有很好的非线性提取和消除噪声的能力,提高了LSSVM的训练能力和泛化能力。通过对比,KPCA比PCA在数据特征提取方面效果更加明显,原因是PCA是一种线性映射方法,对数据处理时忽略了数据之间高于2阶的相互关系;而KPCA是一种非线性映射方法,通过非线性变化来处理数据间的高阶信息,在水文研究领域有很好的应用前景。当然,KPCA在实际应用中也存在不足之处,由于KPCA方法与SVM一样都用到了核函数,而目前对核函数及核参数的选择并无统一标准,关于如何更合理地选择核函数及核参数,有待于进一步深入研究。

[参考文献]

[1] 廖 杰,王文圣,李跃清,等. 支持向量机及其在径流预测中的应用[J]. 四川大学学报:工程科学版,2006,38(6):24-28.
Liao J,Wang W S,Li Y Q,et al. Support vector machine method and its application to prediction of runoff [J]. Journal of Sichuan University:Engineering Science Edition,2006,38(6):24-28. (in Chinese)

[2] Cao L J,Chuan K S,Chong W K. A comparison of PCA,KPCA and ICA for dimensionality reduction in support vector machine [J]. Neurocomputing,2003,55(2):312-336.

[3] 刘全昌,贺国平,张妮娜. KPCA和RSVM结合处理大规模问题研究[J]. 山东科技大学学报:自然科学版,2008,27(1):72-75.
Liu Q C,He G P,Zhang N N. Solutions to the Massive Problems with KPCA and RSVM [J]. Journal of Shandong University of Science and Technology:Natural Science Edition,2008,27(1):72-75. (in Chinese)

[4] 李 岳,温熙森,吕克洪. 基于核主成分分析的铁谱磨粒特征提取方法研究[J]. 国防科技大学学报,2007,29(2):113-116.
Li Y,Wen X S,Lv K H. KPCA-based technique for debris feature extraction [J]. Journal of National University of Defense Technology,2007,29(2): 113-116. (in Chinese)

[5] Twining C J,Taylor C J. The use of Kernel Principal component analysis to model data distributions [J]. Pattern Recognition,2003,36:217-227.

[6] 刘显贵,谢云敏, 陈无畏. 一种基于核主元分析的支持向量机识别方法[J]. 南昌大学学报:理科版,2007,31(2):49-52.
Liu X G,Xie Y M,Chen W W. Research of support vector machine classified method based on kernel principal component analysis [J]. Journal of Nanchang University: Natural Science Edition,2007,31(2):49-52. (in Chinese)

[7] Vapnik V N. Statistical learning theory [M]. New York:Addison Wiley,1998.

[8] Gunn R G. Support vector machines for classification and regression [R]. Southampton: University of Southampton,1998.

[9] 王景雷,吴景社,孙景生,等. 支持向量机在地下水位预报中的应用研究[J]. 水利学报,2003(5):122-128.
Wang J L,Wu J S,Sun J S,et al. Application of support vector machine method in prediction of groundwater level [J]. Journal of Hydraulic,2003(5):122-128. (in Chinese)

[10] 梅 松,程伟平,刘国华. 基于支持向量机的洪水预报模型初探[J]. 中国农村水利水电,2005(3):34-36.
Mei S,Cheng W P,Liu G H. Preliminary discussion on flood forecast model based on support vector machine on flood forecast model based on support vector machine [J]. China Rural Water and Hydropower,2005(3):34-36. (in Chinese)

[11] Tay F E H,Cao L J. Descending support vector machines for financial time series forecasting [J]. Neural Processing Letters,2002,15(2):179-195.

[12] 吴今陪. 基于核函数的主成分分析及应用[J]. 系统工程,2005,23(2):117-120.
Wu J P. Principle component analysis and application based on kernel function [J]. Systems Engineering,2005,23(2):117-120. (in Chinese)

[13] Shawe-Taylor J. Nello cristianini,kernel method for pattern analysis [M]. London:Cambridge University,2004.

[14] 徐 晔,杜文莉,钱 锋. 基于核主元分析和最小二乘向量机的软测量建模[J]. 系统仿真学报,2007,19(17):3873-3875.
Xu Y,Du W L,Qian F. Soft sensor model based on KPCA and Least Square SVM [J]. Journal of System Simulation,2007,19(17): 3873-3875. (in Chinese)

[15] 许先英,丁国栋,高志海,等. 近50年民勤绿洲生态环境演变及综合治理对策[J]. 中国水土保持科学,2006,4(1):40-48.
Xu X Y,Ding G D,Gao Z H,et al. Succession of ecological environment in the last 50 years in Minqin Oasis of Gansu province and its comprehensive controlling countermeasures [J]. Science of Soil and Water Conservation,2006,4(1):40-48. (in Chinese)