EMD 在广西季节降水预报中的应用*

毕硕本[1)] 徐 寅[2)] 覃志年[3)] 陈 譞[1)] 王必强[1)]

[1)](南京信息工程大学计算机与软件学院,南京 210044)

[2)](南京信息工程大学信息与控制学院,南京 210044) [3)](广西壮族自治区气候中心,南宁 530022)

摘 要

气候系统是一种耗散的、具有多个不稳定源的非线性、非平稳系统。该文利用支持向量机(SVM)算法在处理非线性问题中的优越性和经验模态分解(EMD)算法在处理非平稳信号中的优势,采用将 EMD 与 SVM 相结合的短期气候预测方法,并应用到广西季节降水预报中。选取广西 88 个气象观测站 1957—2005 年 6—8 月逐年降水量的距平百分率序列作为试验数据,通过 EMD 算法将标准化处理后的距平百分率序列分解成多个本征模态函数(IMF)分量和一个趋势分量,在分解中针对 EMD 算法存在的端点极值问题选择两种方法分别进行处理,对比得出极值延拓法效果更好。对每个分量构建不同的 SVM 模型进行预测,并通过重构形成最后的预测结果。试验中采用不经 EMD 处理的反向传播(BP)神经网络和 SVM 算法进行对比验证,结果表明:相对于直接预测方法,该文提出的方案均方误差最小,能够较为准确地反映出降水序列未来几年的变化趋势,具有更高的预测精度和较好的推广前景。

关键词:经验模态分解(EMD);支持向量机(SVM);短期气候预测;降水预报;时间序列

引 言

气候系统是一种耗散的高阶非线性系统,在气候预测中,对于处理非线性时序问题有独特优越性的人工神经网络等技术已得到了一定应用。如张迎春等[1]采用基于时间序列的 BP(Back Propagation)神经网络对克拉玛依沙漠进行气温预测分析;金龙等[2]采用了基于遗传算法的 BP 神经网络进行短期气候预测建模研究;陈永义等[3-4]首次将支持向量机方法引入到气象预报试验中,冯汉中等[5]、燕东渭等[6]、刘科峰等[7]在此基础上进行了更深入的探索,取得了较好的预报效果。

同时,气候时间序列也具有典型的非平稳特征[8],可以借助信号处理方法进行平稳化处理,以获得更好的预测效果。Huang 等[9]于 1998 年提出了一种新的信号处理方法——经验模态分解(Empirical Mode Decomposition, EMD)。它将非平稳信号按不同尺度的波动或趋势分解成若干个本征模态函数(Intrinsic Mode Function, IMF)分量及一个趋势分量的线性和。不同的 IMF 分量是平稳信号,具有非线性特征,也具有时间上的局域化特征。经验模式分解结果完全由信号本身决定,是一种自适应信号分解方法,其滤波特性与小波分解非常相似。目前,EMD 算法已在多时间尺度分析[10]、时间序列预测[11]、故障诊断[12]等多个方面获得了较好的应用效果。

本文采用广西夏季的逐年降水距平百分率资料,将 EMD 算法与支持向量机(Support Vector Machine, SVM)时间序列预测算法相结合,进行短期气候预测。经验证,相对于不使用 EMD 的 BP 神经网络和 SVM 进行的预测,本文中提出的方案能够取得更好的效果。

1 经验模态分解

1.1 经验模态分解基本步骤

EMD 算法将时间信号 $X(t)$ 分解成一系列的 IMF 分量,每个 IMF 具有如下两个特征:从全局特性上看,极值点数必须和过零点数一致或者至多相差 1 个;在某一个局部点,极大值包络和极小值包络在该点的值的算术平均值是零。详细分解步骤如

* 中国气象局新技术推广项目(CMATG2009MS19(2))资助。
2009-05-21 收到,2010-04-07 收到再改稿。

下[9]:

(1) 对于输入序列 $X(t)$进行标准化处理,得到序列$S(t)$。判断 $S(t)$的极值数是否大于 2:是,则继续执行;否,则跳转到步骤(4)。

(2) 定义 $H(t)=S(t)$,在 $H(t)$中进行提取 IMF 分量的循环操作,包括:

① 提取 $H(t)$序列中的极大值和极小值。通过插值算法将极大值和极小值插值到整个时间段上,得到 $\max(t)$和 $\min(t)$,并计算算术平均值 $m(t)$。

② 令 $H(t)=H(t)-m(t)$。此时的 $H(t)$有可能为 IMF 分量,通过一系列的约束条件来加以判断:

(Ⅰ) 满足上述的 IMF 分量的两个特征。

(Ⅱ) 为了防止过分迭代,设置 $H(t)$的最大迭代次数为 200。

(Ⅲ) 通过限制两个连续的处理结果之间的标准差 D 的大小进行约束,如式(1)所示。一般 D 的取值在 0.2～0.3 之间:

$$D=\sum_{t=0}^{T}\left[\frac{|H_{k-1}(t)-H_k(t)|^2}{H_{k-1}^2(t)}\right]。\qquad(1)$$

③ 若以上条件都满足,$F_{IM}(t)=H(t)$,跳出循环;否则转到①继续迭代。

(3) 成功提取一个 IMF 分量,并令 $S(t)=S(t)-F_{IM}(t)$,跳转到步骤(1)。

(4) 此时所有的 IMF 分量都被提取出,剩余的 $S(t)$则表现为一个单调或近似单调的趋势项,称为趋势分量 $R_n(t)$。时间序列实现分解:

$$S(t)=\sum_{i=1}^{n}F_{IMi}(t)+R_n(t)。\qquad(2)$$

1.2　经验模态分解关键技术

根据上述的算法原理可以看出,经验模态分解过程中最关键的一步是通过信号的极值点拟合信号包络线,而目前尚未从理论上严格确定采用何种包络线算法[13]。文中采用的是应用最广泛的三次样条插值函数法。三次样条函数需要信号两端数据的一阶或二阶导数作为其边界已知条件,而由 EMD 算法的原理可知,无法直接获得两端点对应的极值,因此许多学者运用多种方法来拟合信号的端点极值,主要有黄大吉等[14]提出的极值延拓法,Zhao 等[15]提出的镜像延拓法,朱金龙等[16]提出的正交多项式拟合法以及邓拥军等[17]提出的线性神经网络法。

针对本文在短期气候预测方向的应用,神经网络延拓方法的速度过慢[18],无法适用实际需求;镜像延拓法可能需截去一部分数据以获得极值点位置,对于本文的短时间序列也不适合。因此将原序列首先进行标准化处理后,分别采用极值延拓法和正交多项式拟合两种方法进行端点延拓,对比结果如图 1 所示。

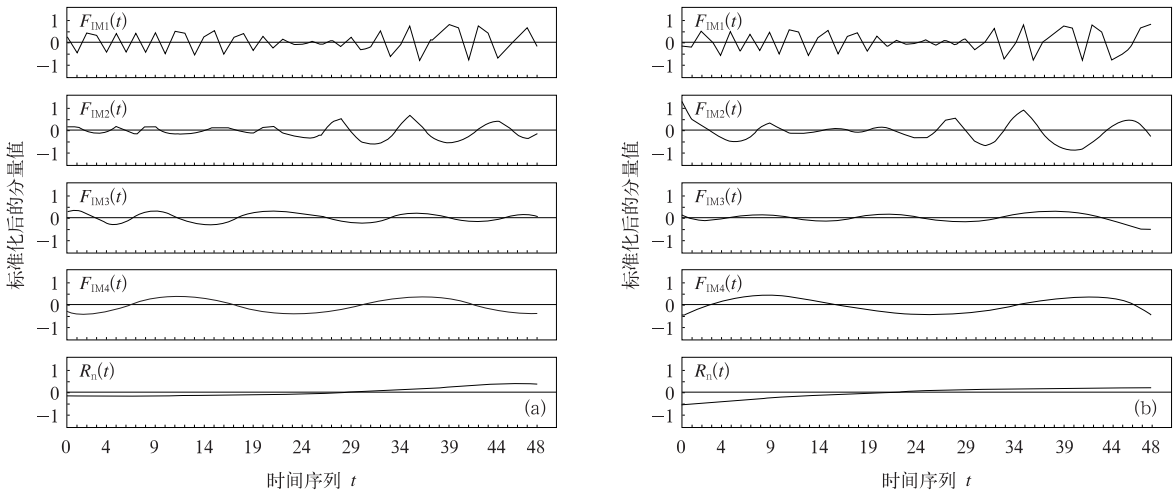

图 1　极值延拓法(a)与正交多项式拟合法(b)的 EMD 处理结果

Fig. 1　Results of EMD based on extrema extending algorithm (a) and orthogonal polynomial fitting algorithm (b)

其中,图 1a 采用的是极值延拓法,图 1b 采用的是正交多项式拟合法。从分解结果可以看出,运用两种端点延拓方法所得分量个数是一致的,均为 4 个 IMF 分量和一个趋势分量;且 $F_{IM1}(t)$的振幅最大,波长最短,之后依次减小,符合 EMD 算法的原理。不过,从图 1b 中可以看出,$F_{IM2}(t)$分量的左端

点处出现了明显的振幅过大，即所谓的端点飞翼现象，这直接影响到了 $F_{IM3}(t)$ 的值。从算法原理上看，正交多项式拟合法[16]是通过构造一个正交多项式函数对离端点最近的有限个已知极值点（由于序列较短，这里取 3 个）进行数据拟合，拟合的标准是最小二乘法。不足之处在于，为了获得最佳的拟合效果，多项式阶数较高，导致在端点处（特别是时间序列的起始零点）振幅过大，曲线严重失真，如图 1b 中的 $F_{IM2}(t)$ 分量；当多项式取一阶或二阶时，曲线会更加平滑，但可能不满足最小二乘的条件；并且所取的拟合极值点个数也是人为确定，没有严格限制。而极值延拓法是根据序列内极值点出现的位置以及与端点的关系，向两侧进行对称的延拓。对延长的新序列使用三次样条插值后，再截去人为延拓的部分。由于使用了已知的极值点作为延拓的端点，从而有效地避免了端点的飞翼现象。因此，实际中采用极值延拓法作为端点极值拟合方法。

2 SVM 时间序列预测

2.1 时间序列构造

设有时间序列$\{x(1),x(2),\cdots,x(N)\}$，根据时间序列的历史数据$\{x(t),x(t-1),\cdots,x(t-m+1)\}$预测未来 $t+k$ 时刻的值 $x(t+k)$，使其满足

$$x(t+k)=F(x(t),x(t-1),\cdots,x(t-m+1)), \quad (3)$$

称为时间序列的预测。而将 SVM 应用于时间序列预测，是指用 SVM 拟合函数 F，将时间序列建模与预测问题转换成 SVM 的回归估计。其中，当 $k=1$ 时，称为单步预测；当 $k>1$ 时，称为多步预测。这种多步预测为迭代式，即每次进行的仍为单步预测，将得到的结果作为下一步预测的输入，来计算接下来的预测值。参数 m 称为嵌入维数，一般根据实际情况选取。

2.2 基于 SVM 的时间序列预测

SVM 是一种基于统计学习理论的机器学习方法，遵循结构风险最小化准则。相对于其他机器学习算法，具有结构简单，全局最优，泛化能力好的特点，更适合解决小样本情况下的学习问题[19]。本文采用由 Suykens 等[20]提出的一种 SVM 改进算法——最小二乘支持向量机（least squares support vector machines，LS-SVM）算法，与传统 SVM 算法相比，可降低计算复杂度，提升求解速度。用于时间序列预测的 LS-SVM 回归算法关键步骤描述如下：

给定一个含 N 个样本的训练集$\{\boldsymbol{x}_k,y_k\}_{k=1}^N$，其中，$\boldsymbol{x}_k$ 为 n 维输入向量，y_k 是一维输出标量。首先通过非线性映射函数 $\varphi(\boldsymbol{x}_k)$ 将输入向量映射到特征空间，并表示成最优化问题：

$$\min J(\boldsymbol{w},\boldsymbol{e})=\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\frac{1}{2}\gamma\sum_{k=1}^{N}e_k^2。 \quad (4)$$

约束条件为：

$$y_k=\boldsymbol{w}^T\varphi(\boldsymbol{x}_k)+b+e_k,\quad k=1,\cdots,N。 \quad (5)$$

式(4)～(5)中，$\boldsymbol{w}$ 为权重向量，e_k 为松弛变量，b 为偏置，γ 是正则化参数，它能够在训练误差和模型复杂度之间取一个折衷以便使所求的函数具有较好的泛化能力。接下来引入拉格朗日函数求解该优化问题：

$$L(\boldsymbol{w},b,\boldsymbol{e},\boldsymbol{a})=J(\boldsymbol{w},\boldsymbol{e})-\sum_{k=1}^{N}\alpha_k[\boldsymbol{w}^T\varphi(\boldsymbol{x}_k)+b+e_k-y_k]。 \quad (6)$$

式(6)中，α_k 为拉格朗日乘子。根据 KKT 优化条件，并定义核函数 $K(\boldsymbol{x},\boldsymbol{x}_k)=\varphi^T(\boldsymbol{x})\varphi(\boldsymbol{x}_k)$，通过求解线性方程组得到回归的决策函数为：

$$f(\boldsymbol{x})=\sum_{k=1}^{N}\alpha_k K(\boldsymbol{x},\boldsymbol{x}_k)+b。 \quad (7)$$

这里，核函数选取常用的径向基核函数：

$$K(\boldsymbol{x},\boldsymbol{x}_k)=\exp\left\{\frac{-\|\boldsymbol{x}-\boldsymbol{x}_k\|}{2\sigma^2}\right\}。 \quad (8)$$

其中，σ 为径向基核宽度，和正则化参数 γ 一样均为待定参数。本文中选用交叉验证法[21]来优化模型参数。

3 基于 EMD 和 SVM 的短期气候预测

本文将经验模态分解与支持向量机的时序预测方法相结合，即首先通过经验模态分解算法将标准化处理后的时间序列分解为多个 IMF 分量和 1 个趋势分量，并对每一个分量分别构造了 1 个 SVM 模型进行预测，再将预测结果线性合成最终的预测序列。

试验中的数据资料来自广西全区 88 个气象观测站 49 年（1957—2005 年）6—8 月逐年降水量距平百分率序列。由于 6—8 月是广西主要的降雨季节，也是广西容易发生洪涝灾害的季节，因此准确地预测降水量的变化趋势有很大的实际意义。介于降水量的逐年波动幅度较大，在使用 EMD 分解前需进

行标准化处理。这里采用 z-score 标准化方法[22]，即

$$V' = \frac{v - \overline{A}}{\sigma_A}。\tag{9}$$

式(9)中，v 为原值，$\overline{A}$ 和 σ_A 分别为序列的均值和标准差，处理后得到标准化的距平百分率序列 $X(t)$。使用 EMD 方法对 $X(t)$进行分解，结果如图 1b 所示。接下来，将分解出的 IMF1～IMF4 分量以及 R_n 趋势分量通过构造不同的 SVM 进行预测。对每个 SVM，根据时序预测方法，选择 $X(t)$中前 40 年的数据，构成训练样本集$\{(x_i, y_i) \mid i = 1, 2, \cdots, 10\}$。其中 x_i 包含 30 年数据，y_i 为对应后一年的值，即

$$\begin{aligned} x_i &= [X(i)X(i+1)\cdots X(i+29)], \\ y_i &= X(i+30), \\ i &= 1, 2, \cdots, 10。 \end{aligned}\tag{10}$$

预测 $X(t)$中后 9 个数据，即 1997—2005 年的降水量，并通过相对误差 e 和均方误差 e_{av}检验预测效果：

$$e = \left| \frac{r_i - f_i}{r_i} \right|,$$

$$e_{av} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left| \frac{r_i - f_i}{r_i} \right|^2}。\tag{11}$$

式(11)中，r_i 为实况值，f_i 为预测值。为了对比验证，使用不经过 EMD 处理的 BP 神经网络和 SVM 算法进行预测。其中，BP 神经网路采用单隐层结构，隐层节点数设为 8，带有可变学习率和动量项系数。首先来看其拟合效果，即用训练好的网络来预测这 10 组样本，结果表明：运用 3 种方案均可达到接近 100%的精度，可见都具有很好的数据拟合能力。接下来进行预测，所得相对误差和相对误差曲线图如表 1 和图 2 所示，预测值与实况值的比较如图 3 所示。

表 1　运用 3 种方案进行预测的相对误差

Table 1　The relative error of three schemes for forecast

方案	1997 年	1998 年	1999 年	2000 年	2001 年	2002 年	2003 年	2004 年	2005 年
BP	0.1830	1.6125	0.1468	0.5132	1.0534	1.2238	0.2323	0.6837	0.4683
SVM	0.1752	1.298	0.0384	0.5728	0.606	1.165	0.0011	0.4136	0.1023
经 EMD 处理的 SVM	0.0270	1.2147	0.0290	0.3656	0.0308	1.044	0.1166	0.4469	0.0600

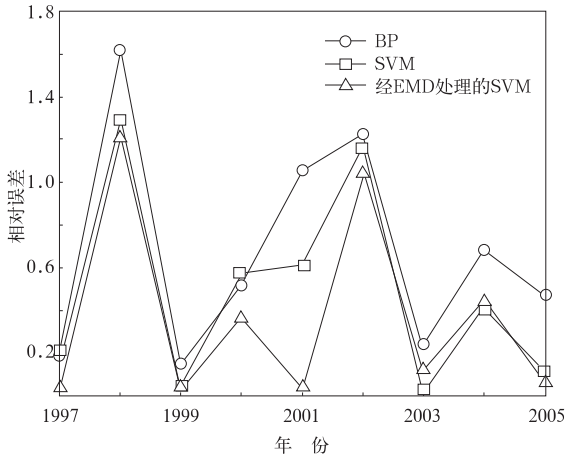

图 2　运用 3 种方案进行预测的相对误差曲线

Fig. 2　The relative error curve of three schemes for forecast

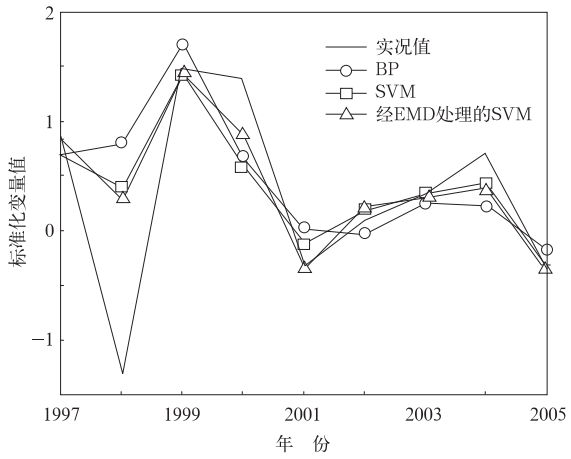

图 3　预测值与实况值的比较

Fig. 3　The comparison between predicted value and actual value

其中，3 种方案的均方误差分别为：0.8344，0.6225和 0.3681。

先比较 SVM 和 BP 神经网络方法。表 1 中，如果只对 1997 年做单步预测，则 BP 神经网络方法的精度近似于 SVM 方法，但随着步长的增加，逐渐反映出了 BP 算法的不足。如在 1998 年和 2001 年，降水量出现了明显的波动，使得 BP 算法的误差均高出其他两种方法 50%左右，在预测曲线上表现为

这一段更加平滑，这暴露了 BP 算法的过学习性。在整体的误差变化中，除了 2000 年，BP 算法产生的误差几乎都是最大，并且在 2005 年处还有增大的趋势。这说明了 BP 算法在处理类似于本文的小样本问题时，泛化能力较弱，而支持向量机算法以结构风险最小化为准则，在样本数量有限的情况下可以更好地反映出变化规律。

经 EMD 处理后的 SVM 方法，从误差曲线可以看出总体处理后趋势变化和原方法比较一致，除 2003 年和 2004 年的误差略高，其余年份都达到了更好的预测效果，从均方根误差上可以更明显看出该方法的优势。说明 EMD 方法将原始序列分解为一系列具有平稳特性的分量，反映出序列在不同时间尺度上的变化规律，更适合于使用机器学习方法进行预测。而在对每个 IMF 分量分别使用 SVM 进行预测时，较大的误差主要集中在 IMF1 和 IMF2 两分量中，说明了由 EMD 处理出的高频分量有时仍带有一定的非平稳性，在今后工作中需要对算法做进一步改善。

4 小 结

在对具有非线性、非平稳特性的气候时间序列进行短期气候预测中，本文采用了经验模态分解方法，对原始序列进行平稳化处理，分解出若干 IMF 分量和一个趋势分量，并分别对每个分量运用支持向量机算法进行预测，将结果重构为最终的预测值。经过广西全区 49 年中的夏季降水量距平百分率数据测试，验证了该方案的优越性，能够更好地反映出降水量的变化趋势，在短期气候预测领域具有较为广泛的应用前景。

参 考 文 献

[1] 张迎春，肖东荣，赵远东. 基于时间序列神经网络的气象预测研究. 武汉理工大学学报，2003，27(2)：237-240.

[2] 金龙，吴建生，林开平，等. 基于遗传算法的神经网络短期气候预测模型. 高原气象，2005，24(6)：981-987.

[3] 陈永义，俞小鼎，高学浩，等. 处理非线性分类和回归问题的一种新方法(Ⅰ)——支持向量机方法简介. 应用气象学报，2004，15(3)：345-354.

[4] 冯汉中，陈永义. 处理非线性分类和回归问题的一种新方法(Ⅱ)——支持向量机方法在天气预报中的应用. 应用气象学报，2004，15(3)：355-365.

[5] 冯汉中，陈永义，成永勤，等. 双流机场低能见度天气预报方法研究. 应用气象学报，2006，17(1)：94-99.

[6] 燕东渭，孙田文，杨艳，等. 支持向量数据描述在西北暴雨预报中的应用试验. 应用气象学报，2007，18(5)：676-681.

[7] 刘科峰，张韧，洪梅，等. 基于最小二乘支持向量机的副热带高压预测模型. 应用气象学报，2009，20(3)：354-359.

[8] 林振山，汪曙光. 近四百年北半球气温变化的分析:EMD 方法的应用. 热带气象学报，2004，24(1)：90-96.

[9] Huang N E，Zheng Shen. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Nonstationary Times Series Aanalysis. Proceedings of the Royal Society，London，1998：903-995.

[10] 黄伟，杨志刚，丁志宏. 基于 EMD 的官厅水库天然年径流量变化多时间尺度分析. 水资源与水工程学报，2008，19(1)：49-52.

[11] 李楠，曾兴雯. 基于 EMD 和神经网络的时间序列预测. 西安邮电学院学报，2007，12(1)：51-54.

[12] 林瑞霖，周平. 基于 EMD 和神经网络的气阀机构故障诊断研究. 海军工程大学学报，2008，20(2)：48-51.

[13] 钟佑明，金涛，秦树人. 希尔伯特-黄变换中的一种新包络线算法. 数据采集与处理，2005，20(1)：13-14.

[14] 黄大吉，赵进平，苏纪兰. 希伯尔特-黄变换的端点延拓. 海洋学报，2003，25(1)：1-11.

[15] Zhao J P，Huang D J. Mirror extending and circular spline function for empirical mode decomposition method. *Journal of Zhejiang University*，2001，2(3)：247-252.

[16] 朱金龙，邱晓晖. 正交多项式拟合在 EMD 算法端点问题中的应用. 计算机工程与应用，2006，23：72-74.

[17] 邓拥军，王伟，钱成春，等. EMD 方法及 Hilbert 变换中边界问题的处理. 科学通报，2001，46(3)：257-263.

[18] 于伟凯. EMD 时频分析方法的理论研究与应用. 秦皇岛：燕山大学，2006.

[19] 杜熊禹. 用于数据挖掘的支持向量机算法研究. 成都：电子科技大学，2007.

[20] Suykens J A K，Lukas L，Vandewalle J. Least squares suppert vecter machine classifiers. *Neural Processing Letters*，1999，9(3)：293-300.

[21] 董春曦，饶鲜，杨绍全，等. 支持向量机参数选择方法研究. 系统工程与电子技术，2004，26(8)：1117-1120.

[22] Han Jiawei，Micheline Kamber. 数据挖掘概念与技术. 范明，孟小峰，译. 北京：机械工业出版社，2008：46.

Application of EMD to Seasonal Precipitation Forecast in Guangxi

Bi Shuoben[1)]　Xu Yin[2)]　Qin Zhinian[3)]　Chen Xuan[1)]　Wang Biqiang[1)]

[1)] (*School of Computer Science and Technology, Nanjing University of Information Science & Technology, Nanjing* 210044)

[2)] (*School of Information and Control Technologies, Nanjing University of Information Science & Technology, Nanjing* 210044)

[3)] (*Climate Center of Guangxi Zhuang Autonomous Region, Nanning* 530022)

Abstract

The climate system is a high order nonlinear system with dissipation. In recent years, the BP neural network algorithm and the Support Vector Machine (SVM) algorithm are applied widely in the short-range climate forecast for its superiority in handling nonlinear time series problem. Besides, the climatic time series are non-stationary, so the signal needs processing to improve its predication result. The Empirical Mode Decomposition (EMD) algorithm introduced by Huang is used to stabilize the climatic time series. Combined with the SVM algorithm, it's used for short-range climate forecast and applied to the seasonal precipitation forecast in Guangxi.

The EMD algorithm decomposes non-stationary signal into several Intrinsic Mode Functions (IMF) components and a remainder with stationary. EMD algorithm doesn't provide a good solution for the endpoints extremes problem, and the extreme extending method is adopted as the endpoints continuation method for short-range climate forecast. Anomaly percentage of accumulated precipitation data are analyzed, which are observed at 88 meteorological observatories in Guangxi from June to August during 1957—2005. Using the EMD algorithm, the time series being standardized are decomposed into four IMF components and a remainder; then a SVM model is built for each component, and the forecasts are composed to the final forecast result. For comparison, BP neural network algorithm and SVM algorithm are adopted to forecast respectively without the EMD algorithm.

Analysis on the predicted values and errors show that, without being processed with EMD, errors of the SVM algorithm are smaller than that of the BP neural network algorithm. So it proves that the generalization capability of BP is weaker than SVM when processing the small sample size problem, whereas SVM algorithm follows the structural risk minimization, and can coincidence the change trend better in condition of finite samples. It shows that the results of the EMD method combined with the SVM algorithm are more accurate. It illustrates that the EMD algorithm can reflect the regularity in different time scales of time series via decomposing into a collection of components with stationarity, which is more suitable for predicting with machine learning methods. The superiority of this scheme makes it widely applicable in precipitation forecast.

Key words: Empirical Mode Decomposition (EMD); Support Vector Machine (SVM); short-range climate forecast; precipitation forecast; time series