

短期负荷多变量混沌时间序列正则化回归局域预测方法

任海军^{1,2} 张晓星¹ 孙才新¹ 文俊浩²

(输配电装备及系统安全与新技术国家重点实验室 重庆 400044)¹

(重庆大学软件学院 重庆 400044)²

摘要 为提高短期负荷预测的精度,提出了多变量混沌时间序列正则化回归局域预测方法。选取对负荷影响程度最大的实感温度因素,构建了多变量时间序列。首先采用互信息法和最小预测误差法确定出时间序列延迟和嵌入维数,并依据确定的重构参数进行短期负荷多变量时间序列的相空间重构,针对局域预测法中邻近点个数少而不能满足最小二乘估计条件的问题,提出了基于正则化回归的多变量时间序列混沌局部预测模型。通过重庆某地区电力系统短期负荷预测的计算实例表明,该方法具有较强的自适应能力和较好的预测效果。

关键词 多变量时间序列,相空间重构,短期负荷预测,正则化回归

中图法分类号 TP301 **文献标识码** A

Regulation Regression Local Forecasting Method of Multivariable Chaotic Time Series in Short-term Electrical Load

REN Hai-jun^{1,2} ZHANG Xiao-xing¹ SUN Cai-xin¹ WEN Jun-hao²

(State Key Laboratory of Power Transmission Equipment & System Security and New Technology, Chongqing 400044, China)¹

(School of Software Engineering, Chongqing University, Chongqing 400044, China)²

Abstract Regulation regression local forecasting method of multivariable chaotic time series in short-term electrical load was proposed, in order to improve the forecasting accuracy of short-term electrical load. The multivariate time series were constructed, by choosing the effective temperature factors with the greatest impact on the load. Firstly, time delay and embedding dimension were confirmed with the methods of mutual information and the minimum predicting error. Secondly, according to reconstruction parameters, the phase space of short-term load multivariate time series was reconstructed. Thirdly, aiming at few neighboring points in the partial predicting method that can not satisfy least-square estimate condition, multivariate time series chaos partial forecasting model based on the regularized regression was presented. Moreover, such model was carried on in practical power load forecast (an electrical power in Chongqing), and the forecasting accuracy was enhanced.

Keywords Multivariate chaotic time series, Phase space reconstruction, Short term load forecasting, Regulation regression

1 引言

混沌时间序列预测是20世纪80年代末发展起来的一种非线性预测方法,已经在天气预报、经济预测、电力负荷预测、股市预测等方面得到了广泛的应用。目前电力负荷预测中,常用的是单变量负荷时间序列的混沌预测方法,即通过短期负荷时间序列建立动态预测模型,只要嵌入维数和延迟时间两个参数选择得合理,单变量时间序列也可以较好地相空间重构,电力负荷预测效果也比较好。但是由于电力短期负荷是一个多维非线性系统,单变量时间序列不能完全重构动力系统的完整信息;同时,由于所获得的时间序列长度有

限,且往往存在噪声,因此单变量时间序列重构的相空间也不能十分准确地描述出动力系统状态变量的演化轨迹。为了能重构出更为准确的相空间,提高短期负荷预测的精确度,国内外提出了基于多变量短期负荷时间序列的混沌预测方法^[1-4]。多变量时间序列包含更加丰富的原动力系统的各种信息,能够重构出更为准确的相空间,从而达到提高预测模型预测精度的目的。

在多变量短期负荷预测中,本文选取对负荷影响程度最大的实感温度因素,构建了多变量时间序列。首先采用互信息法和最小预测误差法确定出时间序列延迟和嵌入维数,并依据确定的重构参数进行短期负荷多变量时间序列的相空间

到稿日期:2009-08-15 返修日期:2009-11-03 本文受国家自然科学基金(基于时空数据挖掘的配电网负荷预测模型及方法研究 No. 5060 7023)资助。

任海军(1978—),男,博士生,讲师,主要研究方向为电力系统负荷预测、地理信息系统, E-mail: jhren@cqu.edu.cn; 张晓星(1972—),男,教授,主要研究方向为负荷预测与局部放电模式识别; 孙才新(1944—),男,教授,中国工程院院士,主要研究方向为输变电安全; 文俊浩(1969—),男,教授,博士生导师,主要研究方向为软件工程。

重构。针对局域预测法中邻近点个数少而不能满足最小二乘估计条件的问题,提出了基于正则化回归的多变量时间序列混沌局域预测模型,并对实际的电力短期负荷进行了预测。目前,混沌时间序列预测方法可分为全局预测法和局域预测法。

全局预测法利用全部的去信息来预测未来值,即用全部已知数据来拟合动力方程。例如用支持向量机建立的全局预测模型,当嵌入维数很高或映射很复杂时,预测精度会迅速下降;同时新的数据加入已知样本时,全局预测模型须重新估计其参数,因此全局预测法计算量较大;并且由于吸引子的结构非常复杂,因此拟合全局动力方程的难度也往往较大。所以全局预测法一般适用于映射不是很复杂,同时噪声干扰比较小的情况。在实际应用中,数据有限,相空间轨迹非常复杂,很难求出真正的映射。

而局域预测法仅利用部分的过去信息来预测未来值,局域动力方程较容易拟合,且计算量较小。Farme 和 Sidorowich 早已证明,在相同的嵌入维数下,局域预测法的效果比全局预测法更好^[5-7]。正是因为如此,局域预测法受到了学术界的重视和关注,成为预测混沌时间序列最基本的方法之一。

2 局域预测法

局域预测法运用较广泛,其基本思想是在相空间中选择有限个预测状态点的最近邻域点,建立一个线性或非线性的局域预报函数进行预测。由于重建相空间本身就是一种非线性的变换,因此即使局域预报函数为线性形式,其本质仍然是一种非线性预报。

对一个 M 维多变量时间序列 $X_1, X_2, \dots, X_N, X_i = (X_{1,i}, X_{2,i}, \dots, X_{M,i}), i=1, 2, \dots, N$, 首先根据相空间重构技术将其嵌入到一个 m 维空间,构造出 m 维相空间的轨迹序列:

$$\begin{aligned} V_{J_0} &= (x_{1,J_0}, x_{1,J_0-\tau_1}, \dots, x_{1,J_0-(m_1-1)\tau_1}, \dots, x_{M,J_0}, \\ &\quad x_{M,J_0-\tau_M}, \dots, x_{M,J_0-(m_M-1)\tau_M}) \\ V_{J_0+1} &= (x_{1,J_0+1}, x_{1,J_0+1-\tau_1}, \dots, x_{1,J_0+1-(m_1-1)\tau_1}, \dots, \\ &\quad x_{M,J_0+1}, x_{M,J_0+1-\tau_M}, \dots, x_{M,J_0+1-(m_M-1)\tau_M}) \quad (1) \\ V_N &= (x_{1,N}, x_{1,N-\tau_1}, \dots, x_{1,N-(m_1-1)\tau_1}, \dots, x_{M,N}, x_{M,N-\tau_M}, \\ &\quad \dots, x_{M,N-(m_M-1)\tau_M}) \end{aligned}$$

式中, $J_0 = \max_{1 \leq i \leq M} (m_i - 1)\tau_i + 1$ 。

根据 Takens 定理,只要嵌入维数 $m(m=m_1+m_2+\dots+m_M)$ 和延迟时间 τ_i 选择合理,重构相空间在嵌入空间的轨迹就与微分同胚意义下的原系统动力学等价,因此存在光滑映射 $f: R_m \rightarrow R_m$, 使得:

$$V_{n+1} = f(V_n)$$

在实际中,不可能真正求得 f , 只能根据有限的观测数据构造映射 $\tilde{f}$ 来近似逼近,并满足:

$$\sum_{n=J_0}^N (x_{1,n+1} - \tilde{f}(x_{1,n}, x_{1,n-\tau_1}, \dots, x_{1,n-(m_1-1)\tau_1}, \dots, x_{M,n},$$

$$A = \begin{bmatrix} k & \sum_{i=1}^k x_{1,i} & \sum_{i=1}^k x_{1,i-\tau_1} & \dots & \sum_{i=1}^k x_{M,i-(m_M-1)\tau_M} \\ \sum_{i=1}^k x_{1,i} & \sum_{i=1}^k x_{1,i}^2 & \sum_{i=1}^k x_{1,i-\tau_1} x_{1,i} & \dots & \sum_{i=1}^k x_{M,i-(m_M-1)\tau_M} x_{1,i} \\ \sum_{i=1}^k x_{1,i-\tau_1} & \sum_{i=1}^k x_{1,i} x_{1,i-\tau_1} & \sum_{i=1}^k x_{1,i-\tau_1}^2 & \dots & \sum_{i=1}^k x_{M,i-(m_M-1)\tau_M} x_{1,i-\tau_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^k x_{M,i-(m_M-1)\tau_M} & \sum_{i=1}^k x_{M,i-(m_M-1)\tau_M} x_{1,i} & \sum_{i=1}^k x_{M,i-(m_M-1)\tau_M} x_{1,i-\tau_1} & \dots & \sum_{i=1}^k x_{M,i-(m_M-1)\tau_M}^2 \end{bmatrix}$$

$$x_{M,n-\tau_M}, \dots, x_{M,n-(m_M-1)\tau_M}))^2.$$

最小,其中 $J_0 = \max_{1 \leq i \leq M} (m_i - 1)\tau_i + 1$ 。

在局域预测中, $\tilde{f}$ 的函数逼近形式通常采用一阶线性方程,即:

$$\begin{aligned} x_{1,n+1} &= a + \sum_{l=1}^M \sum_{j=1}^{m_l} b_{l,j} x_{l,n-(j-1)\tau_l} \\ &= a + b_{1,1} x_{1,n} + \dots + b_{1,m_1} x_{1,n-(m_1-1)\tau_1} + \dots + b_{M,m_M} \\ &\quad x_{M,n-(m_M-1)\tau_M} \end{aligned} \quad (2)$$

式中, $a, b_{1,1}, b_{1,2}, \dots, b_{M,m_M}$ 为拟合参数。

因此,对于方程式(1),若已知拟合参数 $a, b_{1,1}, b_{1,2}, \dots, b_{M,m_M}$, 便可实现 $x_{1,n+1}$ 的预测,而拟合参数的求取常采用最近邻域法,即利用欧氏距离求得预测状态点的 k 个最近邻域点,由最近邻域点及其一步演化轨迹点通过最小二乘法对参数 $a, b_{1,1}, b_{1,2}, \dots, b_{M,m_M}$ 进行估计。

设需要预测量为 $x_{1,n+1}$, 对应相空间相点为 V_{n+1} , 预测中心点为 V_n , 经计算得到 V_n 的最近邻域点为 $V_i, i=1, 2, \dots, k$, 相点 V_i 的一步演化点为 V_{i+1} , 即存在:

$$V_1 \rightarrow x_{1,2}$$

$$V_2 \rightarrow x_{1,3}$$

$\vdots$

$$V_k \rightarrow x_{1,k+1}$$

将以上最近邻域点对应量代入方程(1),有:

$$x_{1,i+1} = a + \sum_{l=1}^M \sum_{j=1}^{m_l} b_{l,j} x_{l,i-(j-1)\tau_l} \quad i=1, 2, \dots, k \quad (3)$$

由最小二乘法,使得:

$$\sum_{i=1}^k (x_{1,i+1} - a - \sum_{l=1}^M \sum_{j=1}^{m_l} b_{l,j} x_{l,i-(j-1)\tau_l})^2 \quad (4)$$

最小。分别对式(2)中 $a, b_{1,1}, b_{1,2}, \dots, b_{M,m_M}$ 求偏导有:

$$\begin{aligned} \sum_{i=1}^k (x_{1,i+1} - a - \sum_{l=1}^M \sum_{j=1}^{m_l} b_{l,j} x_{l,i-(j-1)\tau_l}) &= 0 \\ \sum_{i=1}^k (x_{1,i+1} - a - \sum_{l=1}^M \sum_{j=1}^{m_l} b_{l,j} x_{l,i-(j-1)\tau_l}) x_{1,i} &= 0 \end{aligned} \quad (5)$$

$\vdots$

$$\sum_{i=1}^k (x_{1,i+1} - a - \sum_{l=1}^M \sum_{j=1}^{m_l} b_{l,j} x_{l,i-(j-1)\tau_l}) x_{1,i-(m_M-1)\tau_M} = 0$$

写成矩阵形式为: $C=AB$

其中

$$C = \begin{bmatrix} \sum_{i=1}^k x_{1,i+1} \\ \sum_{i=1}^k x_{1,i+1} x_{1,i} \\ \sum_{i=1}^k x_{1,i+1} x_{1,i-\tau_1} \\ \vdots \\ \sum_{i=1}^k x_{1,i+1} x_{M,i-(m_M-1)\tau_M} \end{bmatrix}, B = \begin{bmatrix} a \\ b_{1,1} \\ b_{1,2} \\ \vdots \\ b_{M,m_M} \end{bmatrix}$$

因此,线性回归方程的拟合参数为:

$$\hat{B} = (\hat{a}, \hat{b}_{1,1}, \hat{b}_{1,2}, \dots, \hat{b}_{M,m_M})^T = (A^T A)^{-1} \quad (6)$$

式中,上标 T 表示转置, -1 表示矩阵的逆。当 A 不是列满秩或者当 A 的条件数过大时,矩阵 $A^T A$ 就开始变得接近奇异,此时会导致参数估计结果变得不可信。

由 $\hat{B}$ 便可实现对 $X_{1,n+1}$ 的预测,即预测值 $\hat{x}_{1,n+1}$ 为:

$$\hat{x}_{1,n+1} = \hat{a} + \sum_{l=1}^M \sum_{j=1}^{m_l} \hat{b}_{l,j} x_{l,n-(j-1)} \tau_l \quad (7)$$

局域预测方法认为相空间中某一点的演化行为可由其邻近点的演化行为反映出来,即在相空间中寻找与预测点的若干最邻近的点,并通过邻域点建立起一个线性回归局域预测模型来达到对未来状态预测的目的。该方法计算简单可行,而且只要能够实时地拟合局部模型的系数,就能够较精确地反映整个吸引子的演化特性,但是这种方法只反映吸引子的局部特性,一旦超出相应的区域范围,局部模型的精度便会下降,甚至完全失效。

3 正则化回归局域预测法

在局域预测模型中有两个比较关键的问题需要解决,即模型参数的估计和邻域点数 k 的选取。对于模型参数估计主要采用上述的最小二乘法;而对于邻域点数 k 的选取,目前还没有确定统一的选取方法,一般来讲,在局部线性预测中主要采用的是文献[4]中提出的解决方法,该方法要求邻域点 k 在 $2m+1$ 至 $2m+10$ 中取值,即要求 $k \gg m$ 。但由于实际问题中样本总数总是有限而不能满足要求,因此相空间中某些点就找不到足够数目的邻近点,条件 $k \gg m$ 满足不了。对于该问题出现如果当 k 接近甚至达到 $k < m$ 的时候,模型中数据矩阵列向量就容易出现近似线性相关性,也即数据矩阵列向量间就会存在多重共线性,最小二乘估计就会变得非常不稳定,从而影响预测结果的精度和可信度。此时,回归方程需要改进,如采用正则化估计进行改善^[8,9]。

由奇异值分解定理可知,任何一个矩阵 $A \in R^{k,m}$,都存在矩阵 $P = [p_1, p_2, \dots, p_r] \in R^{k,r}$ 和 $Q = [q_1, q_2, \dots, q_r] \in R^{m,r}$,使 $A = P\xi Q^T$ 。其中 $P^T P = I, Q^T Q = I, \xi = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$ 。即:

$$A = \sum_{i=1}^r \lambda_i P_i Q_i^T \quad (8)$$

对于线性回归模型,将矩阵 A 的奇异值分解 $A = P\xi Q^T$ 带入到最小二乘估计 $\hat{B} = (A^T A)^{-1} A^T C$ 中,可得到:

$$\hat{B} = Q\xi^{-1} P^T C = \sum_{i=1}^r \frac{(p_i, C)}{\lambda_i} q_i \quad (9)$$

则最小二乘估计的均方误差为:

$$\text{MSE}(\hat{B}) = \sigma^2 \sum_{i=1}^r \frac{1}{\lambda_i} \quad (10)$$

式中, σ^2 为 a 的方差。

因此,当矩阵 A 中的列向量之间存在多重共线性时,矩阵 A 中就会出现至少一个奇异值 λ_i 非常小,相对来讲 $1/\lambda_i$ 将会变得非常大,这就必然导致 $\hat{B}$ 的均方误差特别大,因此 $\hat{B}$ 估计就会变坏。由于矩阵 A 中的较小的奇异值 λ_i 对最小二乘估计算法的影响太大,为了保持最小二乘估计算法稳健性,应该降低这些方向的权值。所以必须对最小二乘估计作修改,得到如下正则化估计:

$$\hat{B} = \sum_{i=1}^r w_i \frac{(p_i, C)}{\lambda_i} q_i \quad (11)$$

上式在每个奇异值代表的方向引入了一个权值 w_i 。若 λ_i 较大,对这样的方向令 $w_i = 1$;若 λ_i 较小,对这样的方向就选择较小的 w_i 。小的权值可以消除小奇异值的影响,从而最小二乘估计的均方误差就会减少。改进的方法有很多,以下介绍两种常用的正则化估计方法:

1) 从减少均方误差的角度出发,引入岭估计(Ridge Regression,简称 RR);

2) 从消除 X 的列向量间的多重共线性关系出发,引入主成分估计(Principal Components Regression,简称 PCR)。

① 岭估计

岭估计是对常规的回归方程的正则化改进,其定义如下:

定义 1 设 $k \geq 0$,称

$$\hat{\beta}(k) = (X^T X + \mu I)^{-1} X^T Y \quad (12)$$

为 β 的岭估计,其中 μ 为岭参数, I 为单位矩阵,由岭估计建立的回归方程称为岭回归。当 $\mu = 0$ 时, $\hat{\beta}(0) = (X^T X)^{-1} X^T Y$ 即为 β 的最小二乘估计结果。

最小二乘估计是对 $\|C - AB\|_2$ 极小化得到的,在出现多重共线性时,某些 λ_i 就会非常小,所以 $\hat{B}$ 的范数就会非常大,为避免出现这种情况,引入岭估计来处理极小化问题:

$$\|C - AB\|_2 + \mu \|B\|_2 \quad (13)$$

此时 μ 是一个正的常数,这样做的主要目的是处罚大范数 $\hat{B}$,可以让解的范数变小。由上式就可以得到岭估计为:

$$\hat{B}_{RR} = (A^T A + \mu I)^{-1} A^T C \quad (14)$$

将 A 的奇异值分解 $A = P\xi Q^T$ 代入上式,得:

$$\hat{B}_{RR} = (Q\xi^2 Q^T + \mu I)^{-1} Q\xi P^T C = \sum_{i=1}^r \frac{\lambda_i}{\lambda_i^2 + \mu} (p_i, C) q_i \quad (15)$$

显然,RR 估计的过滤因子为 $w_i = \frac{\lambda_i^2}{\lambda_i^2 + \mu}$;

对于 RR 估计,参数 μ 的选取是十分重要的问题,近年来提出了许多确定 μ 的原则和方法,各有优缺点。本文中,对 μ 取作残差方差的估计 σ^2 ,残差方差的估计为:

$$\sigma^2 = \frac{1}{k-r} \|C - A\hat{B}\|_2^2 = \frac{1}{k-r} \sum_{i=r+1}^k (p_i^T, C)^2 \quad (16)$$

② 主成分估计

由于回归方程中会出现一些最小奇异值使得最小二乘估计方差变大,为了降低这种奇异值影响,可以从最小二乘估计表达式中直接去掉最小奇异值对应的估计,也就是在式(9)的和式中截掉第 l 项之后的项对应的估计即 PCR 估计,记作:

$$\hat{B}_{PCR} = \sum_{i=1}^l \frac{(p_i, C)}{\lambda_i} q_i \quad (17)$$

数 l 在主成分估计中称作因子数。显然,过滤因子

$$w_i = \begin{cases} 1, & 1 \leq i \leq l \\ 0, & l+1 \leq i \leq r \end{cases}$$

对非线性动力系统,如果系统中吸引子的拓扑维数 m 能够估计出来,则因子数 l 可取为固定值:

$$l = [m] \quad (18)$$

4 正则化回归局域预测法算法流程

基于正则化回归多变量混沌时间序列局部线性预测模型主要针对多变量混沌时间序列局部线性预测线性回归模型,

利用正则化估计中的岭估计或主成分估计算法进行模型参数 $\hat{A}$ 的估计,再利用式(7)进行预测。预测步骤如下:

- ① 重构参数的确定:针对负荷和实感温度时间序列,采用互信息法和最小预测误差法确定时间延迟 τ_i 和嵌入维数 m_i ;
- ② 重构相空间:根据延迟时间 τ_i 和嵌入维数 m_i 重构出多变量的相空间;
- ③ 给定一个合适的邻近点数目 k ,再找出与当前点最近的 k 个邻近点;
- ④ 确定出正则化估计参数、岭估计中 μ 和主成分估计中 l 的值;
- ⑤ 根据式(15)及式(17)分别做出 $\hat{A}$ 的估计;
- ⑥ 根据式(7)做出预测;
- ⑦ 对于新的目标点,重复步骤①至步骤⑥直到结束。

5 短期电力负荷的多变量混沌局域预测法

在短期负荷预测中,除历史负荷时间序列外,影响因素中的气象因素序列可从气象台获得。通过分析发现,在多变量时间序列预测中不同因素的选取对预测结果会有不同的影响,一般根据影响因素对负荷分量变化的影响程度大小来选择部分因素变量,选取那些对动力系统影响较大的因素更能充分地揭示这一系统的动力学特征,可以获得更为理想的预测效果。

为简化分析,本文首先考虑对负荷影响最大的实感温度因素。温度的上升和下降对负荷的影响特别敏感,尤其是夏季和冬季,而湿度和风力等因素的变化在一定程度上也对负荷有较大的影响。实感温度体现了气温、湿度和风速条件的综合影响。因此,本文使用负荷和实感温度两个时间序列对短期负荷进行预测。

以重庆市某电网2008年1月1日到2008年3月31日的每天24点负荷数据和温度数据为依据,对以上方法进行仿真分析,其负荷时间序列为 $X_1, X_2, \dots, X_{2184}$,其中 $X_i = (X_{1,i}, X_{2,i}), i=1, 2, \dots, 2184$, $X_{1,i}$ 为负荷时间序列, $X_{2,i}$ 为实感温度时间序列,则相空间重构为:

$$V_n = (x_{1,n}, x_{1,n-\tau_1}, \dots, x_{1,n-(m_1-1)\tau_1}, x_{2,n}, x_{2,n-\tau_2}, \dots, x_{2,n-(m_2-1)\tau_2})$$
 (19)

式中, $n = \max_{i \in \{1,2\}} (m_i - 1) \tau_i + 1, \dots, 2184$; τ_i 和 m_i 为延迟时间和嵌入维数, $i=1, 2$ 。

对多变量时间序列延迟时间的计算采用互信息法^[10],分别对电力负荷时间序列实感温度影响因素时间序列和进行了分析计算,得到负荷时间序列延迟时间为7,实感温度时间序列延迟时间为2,即 $\tau_1=7, \tau_2=2$ 。

根据最小预测误差法的嵌入维数求取方法^[11-15],设负荷序列和温度序列的最大嵌入维数均为10,计算不同维数对应的平均一步绝对预测误差,分析发现最小绝对预测误差对应的最优嵌入维数分别为负荷时间序列嵌入维数 $m_1=10$ 、实感温度序列嵌入维数 $m_2=3$,则多变量混沌系统嵌入维数 $m=m_1+m_2=13$ 。

5.1 与常规算法比较

将2008年1月至3月每天24点负荷数据作为局域预测方法的模型建立和参数辨识的基础,对该网2008年4月1日

24点负荷进行预测。

为了测试本文提出算法的性能,选取邻域点数目为 $k=16>13=m_1+m_2$,即找出与当前点最近的16个邻域点。主成分估计中 $l=\lceil m_1+m_2 \rceil=13$ 。按照1.3节所述正则化回归局域预测法算法流程进行计算。将正则化回归模型中的岭估计和主成分估计的预测结果与常规多变量时间序列预测结果进行比较,预测结果如表1所列。

表1 2008年4月1日24点负荷预测结果

序号	实际负荷 万 kW	常规算法	误差	岭估计	误差	主成分	误差
1	29.33	28.66	-2.29	29.56	0.79	29.53	0.69
2	27.26	28.56	4.75	27.52	-0.96	27.57	-1.13
3	25.74	25.76	0.08	25.84	0.38	25.88	0.54
4	22.91	24.13	5.31	23.59	-2.95	23.54	-2.75
5	23.05	23.82	3.32	23.43	-1.67	23.34	-1.26
6	22.24	22.88	2.87	22.51	1.23	22.49	1.13
7	23.53	24.24	3.01	24.13	-2.55	23.99	-1.96
8	26.76	25.68	-4.05	27.20	-1.65	27.23	-1.75
9	31.48	30.76	-2.29	32.32	2.68	32.29	2.58
10	41.95	44.30	5.61	43.57	3.87	43.38	3.42
11	47.37	45.40	-4.16	48.48	2.35	48.46	2.31
12	50.42	53.57	6.25	51.76	-2.65	51.41	-1.96
13	47.41	49.54	4.49	48.21	-1.68	48.68	-2.68
14	46.43	49.98	7.65	47.86	3.07	47.66	2.64
15	48.06	45.70	-4.91	48.39	-0.68	48.69	1.32
16	49.90	52.07	4.36	50.33	0.87	50.56	-1.32
17	50.09	52.84	5.49	50.67	1.15	50.41	0.64
18	49.56	52.88	6.70	50.87	2.64	50.54	1.98
19	46.87	44.73	-4.57	47.87	2.13	47.97	2.35
20	45.16	47.76	5.76	46.23	2.36	45.77	1.34
21	48.35	51.56	6.65	48.86	-1.06	48.66	-0.64
22	46.83	45.57	-2.69	47.93	2.34	47.70	1.85
23	41.34	42.02	1.65	42.20	2.09	42.32	2.36
24	36.11	36.76	1.80	36.71	1.67	36.55	1.23

表1中,常规算法的最大误差为7.65%,岭估计的最大预测误差为3.87%,主成分估计的最大误差为3.42%,其平均预测误差分别为4.2%、1.89%和1.74%,通过对二者的比较结果显示,当邻域点数量不满足 $k \gg m_1+m_2$ 条件时,常规算法的误差较大,而岭估计和正则化算法则能保持较好的预测精度。

5.2 不同临近点数量选取误差的比较

采用不同邻域点数 k ,分别利用正则化回归模型中的岭估计和主成分估计算法、常规多变量时间序列算法对4月1日24点负荷进行预测,预测结果如表2所列。

表2 不同邻域点数量下3种算法预测结果比较

算法	k				
	26	20	16	14	13
常规	1.51	1.85	4.2	8.9	—
岭估计	1.49	1.74	1.89	—	—
主成分估计	1.58	1.65	1.74	2.04	2.52

表2的预测结果表明当邻点数远大于嵌入维数时,3种算法的预测误差均较小。当邻域点数逐渐减小时,常规算法的误差增大,岭估计的误差也有所增加,而主成分估计的误差比较稳定,变化不大。当邻域点数接近甚至等于嵌入维数时,常规算法的误差增幅最大,已经失效,而正则化法中的主成分估计仍有较小的预测误差。

结束语 本文选取对负荷影响程度最大的实感温度因素,构建了多变量时间序列,进行了电力负荷的多时间序列相

空间重构,并针对每一维时间序列采用互信息法进行延迟时间的确定,最优嵌入维数的取值方法采用最小预测误差法。根据重构相空间混沌吸引子结构,建立了多变量时间序列的局域预测模型。针对混沌局域预测法中邻近点个数少而不能满足最小二乘估计条件的问题,从减少均方误差的角度出发,引入岭估计;从消除 X 的列向量间的多重共线性关系出发,引入主成分估计,进而提出了基于正则化回归的多变量时间序列混沌局域预测模型及方法,并对实际的电力短期负荷进行了预测。实例仿真结果表明,该预测方法预测精度高,优于传统的混沌局域预测法,在邻域点数接近甚至等于嵌入维数时,预测精度仍然较佳。

参 考 文 献

[1] Stathopoulos A, Karlaftis M G. A Multivariate State Space Approach for Urban Traffic Flow Modeling and Prediction[J]. Transportation research part C, 2003, 11(2): 121-135

[2] Reick C H, Page B. Time series prediction by multivariate next neighbor methods with application to zooplankton forecasts[J]. Mathematics and Computers in Simulation, 2000, 52: 289-310

[3] Jayawardena A W, Li W K, Xu P. Neighborhood Selection for local modeling and prediction of hydrological time series[J]. Journal of Hydrology, 2002, 258(1-4): 40-57

[4] 雷绍兰. 基于电力负荷时间序列混沌特性的短期负荷预测方法

研究[D]. 重庆:重庆大学, 2005

[5] 方仍存, 周建中, 彭兵, 等. 电力负荷混沌动力特性及其短期预测[J]. 电网技术, 2008, 32(4): 61-65

[6] 杜杰, 陆金桂, 曹一家. 杜短期负荷预测最大 Lyapunov 指数预报模式预测值的判定[J]. 电网技术, 2006, 30(20): 20-24

[7] Farmer J D, Sidorowich J J. Predicting chaotic time series[J]. Phys. Rev. Lett., 1987, 59(8): 845-848

[8] 方芬. 多变量混沌时序正则化局部线性预测[J]. 金陵科技学院学报, 2007, 32(2): 9-12

[9] 赵敏, 樊印海, 孙辉. 电力推进船舶电力负荷的多变量混沌局域预测[J]. 系统仿真学报, 2008, 20(11): 2797-2799

[10] 吕小青, 曹彪, 曾敏, 等. 确定延迟时间互信息法的一种算法[J]. 计算物理, 2006, 23(2): 184-188

[11] 蒋传文, 袁智强, 侯志俭, 等. 高嵌入维混沌负荷序列预测方法研究[J]. 电网技术, 2004, 28(3): 25-28

[12] 王海燕, 盛昭瀚, 张进. 多变量时间序列复杂系统的相空间重构[J]. 东南大学学报: 自然科学版, 2003, 33(1): 115-118

[13] Cao Liangyue, Mees A, Judd K. Dynamics from Multivariate Time Series[J]. Physica D, 1998, 121: 75-88

[14] Porporato A, Ridolfi L. Multivariate Nonlinear Prediction of River Flows[J]. Journal of Hydrology, 2001, 248: 109-122

[15] Yang Hongming, Duan Xianzhong. Chaotic Characteristics of Electricity Price and its Forecasting Model[C] // IEEE CCECE 2003. Montreal, 2003: 659-662

(上接第 219 页)

表 4 各种检测算法检测结果对比分析

种类	正确率(%)				
	SCG-BP	弹性 BP	LM-BP	SVM	集成 ANN 和 SVM
正常	92.2	95.75	99.64	98.42	99.71
Probe	62.7	92.71	92.71	98.57	99.85
DoS	99.9	97.47	95.98	98.77	99.97
U2L	0.00	48.00	0.00	64.00	76.00
R2L	56.9	95.73	97.75	97.33	100.00
其他	0.00	38.65	50.24	71.50	86.47
总计	97.29	96.67	96.38	98.51	99.82

Software Engineering, 1987, 13(2): 222-232

[2] 杨种学, 杨宁. 基于 BP 神经网络的异常入侵检测方法[J]. 计算机, 2006, 6: 42-45

[3] 陈刚, 秦茗. 基于数据挖掘的入侵检测研究[J]. 计算机仿真, 2006, 22(5): 43-55

[4] 王正群, 陈世福, 陈兆乾. 并行学习神经网络集成方法[J]. 计算机学报, 2005, 28(3): 402-408

[5] 饶鲜, 董春曦, 杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报, 2003, 14(4): 798-803

[6] <http://kdd.ics.uci.edu/databases/kddcup99/task.htm>, 2007-11-15

[7] Komorowski J O. ROSETTA—A rough set toolkit for analysis of data[C] // Fifth International Workshop on Rough Sets and Soft Computing. Tokyo, Japan, 1997: 403-407

[8] Sung A. Ranking Importance of Input Parameters of Neural Networks[J]. Expert Systems with Applications, 15: 405-41

[9] Burges C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167

[10] McClelland R. Parallel Distributed Processing: Explorations in the Microstructure of Cognition [M]. Cambridge, MA: MIT Press, Vol 1, 1986

[11] Schapire R E. The boosting approach to machine learning: An overview[C] // MSRI Workshop on Nonlinear Estimation and Classification. 2002

[12] Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12: 993-1001

[13] 齐德昱, 葛超, 葛初. 混合核支持向量回归及对社会用电量的预测[J]. 重庆工学院学报: 自然科学版, 2009, 23(10): 50-52

综合以上的结果,采用 ANN 和 SVM 的网络集成具有如下优点:(1)极大地提高了集成网络的泛化能力。通过集成,对各类攻击的检测率都有显著提高。(2)在各分类器均缺乏攻击全部知识的情况下,通过相互协作与知识互补,集成学习,对小样本以及未知的攻击类型都有较高的识别率。

结束语 BP 算法具有对不确定性的学习与适应能力,能很好地满足入侵检测的要求,但容易陷入局部最小,训练时间长,所以本文引入 3 种算法对 BP 算法进行改进;支持向量机在解决小样本、非线性及高维问题时具有明显优势,然而对大数据集来说,支持向量机的时空耗费非常大。综上所述,单一的学习算法往往有各自的缺陷,限制了其在入侵检测中的应用,因此本文将这些具有优良分类性能的检测器进行集成。研究表明,在集成学习中,各子分类器差异度较大时才有较好的分类正确性。我们用同一训练集对不同子分类器进行重复训练来实现集成个体的差异度,在各分类器均缺乏攻击全部知识的情况下,通过相互协作与知识互补,集成学习,来达到较高的检测率和对未知攻击类型的识别。

参 考 文 献

[1] Enning D D. An Intrusion Detection Model[J]. IEEE Trans on

• 224 •