一种基于 L_1 范数正则化的回声状态网络

韩 敏 [1] 任伟杰 [1] 许美玲 [1]

摘 要 针对回声状态网络存在的病态解以及模型规模控制问题, 本文提出一种基于 L_1 范数正则化的改进回声状态网络. 该方法通过在目标函数中添加 L_1 范数惩罚项, 提高模型求解的数值稳定性, 同时借助于 L_1 范数正则化的特征选择能力, 控制网络的复杂程度, 防止出现过拟合. 对于 L_1 范数正则化的求解, 采用最小角回归算法计算正则化路径, 通过贝叶斯信息准则进行模型选择, 避免估计正则化参数. 将模型应用于人造数据和实际数据的时间序列预测中, 仿真结果证明了本文方法的有效性和实用性.

关键词 回声状态网络, 正则化, 最小角回归, 信息准则, 多元时间序列

引用格式 韩敏, 任伟杰, 许美玲. 一种基于 L_1 范数正则化的回声状态网络. 自动化学报, 2014, **40**(11): 2428−2435

DOI 10.3724/SP.J.1004.2014.02428

An Improved Echo State Network via L_1-Norm Regularization

HAN Min[1] REN Wei-Jie[1] XU Mei-Ling[1]

Abstract Considering the ill-posed problem and the model scale control of echo state network, an improved echo state network based on L_1-norm regularization is proposed. In order to improve the numerical stability, the proposed method adds an L_1-norm penalty term in the objective function. Meanwhile, the method can also control the complexity of the network and prevent overfitting by using feature selection capability of L_1-norm regularization. To solve the L_1-norm regularization model, we adopt the least angle regression algorithm to calculate regularization path and select suitable model through Bayesian information criterion, which can avoid the estimations of regularization parameter. The model is applied to the time series predictions of both synthetic dataset and practical dataset. The simulation results show the effectiveness and practicality of the proposed method.

Key words Echo state network (ESN), regularization, least angle regression (LARS), information criterion, multivariate time series

Citation Han Min, Ren Wei-Jie, Xu Mei-Ling. An improved echo state network via L_1-norm regularization. *Acta Automatica Sinica*, 2014, **40**(11): 2428−2435

传统递归神经网络通常采用基于梯度的学习算法, 训练过程中普遍存在收敛速度慢、易陷入局部最优的问题, 使其在实际应用中受到很大限制. 回声状态网络 (Echo state networks, ESN)[1] 是一种新型的递归神经网络, 与传统的递归神经网络相比, ESN 网络的稳定性更好, 而且训练过程简单、速度快, 只需要训练网络的输出权值. 此外, ESN 网络具有良好的非线性映射能力, 在非线性预测等领域取得了良好的结果, 得到学者的广泛关注[2−4].

ESN 网络虽然具有其他递归神经网络无法比拟的优点, 但是在网络结构和学习算法等方面仍不够成熟[5]. ESN 网络的核心结构为储备池, 储备池随机生成且固定不变, 包含大量稀疏连接的神经元, 神经元个数在几百到几千之间. 储备池的规模过小, 容易出现欠拟合现象, 则无法体现网络强大的非线性映射能力; 储备池的规模过大, 则会引入大量的冗余特征和无关特征, 容易出现过拟合现象, 进而影响网络的泛化性能, 并伴随着较大的计算复杂度. 因此, 针对实际问题, 有效控制储备池的规模, 建立最优的网络结构, 是亟待解决的关键问题[6].

为了提高网络的预测性能, 学者提出采用剪枝方法控制模型规模[7], 以提高网络的泛化性能. 借助于剪枝算法, 可以完全移除对预测变量影响不大的冗余特征和无关特征, 等价于模型的特征选择过程. 文献 [8] 采用后向选择、最小角回归等子集选择方法, 剔除 ESN 网络的无关节点, 达到控制模型规模的目的. 文献 [9] 根据多元响应稀疏回归算法对神经元进行排序, 采用留一法选择网络结构, 提出最优剪

收稿日期 2013-11-06 录用日期 2014-05-02
Manuscript received November 6, 2013; accepted May 2, 2014
国家重点基础研究发展计划 (973 计划) (2013CB430403), 国家自然科学基金 (61374154) 资助
Supported by National Basic Research Program of China (973 Program) (2013CB430403) and National Natural Science Foundation of China (61374154)
本文责任编委 王占山
Recommended by Associate Editor WANG Zhan-Shan
1. 大连理工大学电子信息与电气工程学部 大连 116024
1. Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024

枝极限学习机. 子集选择方法能够实现模型的规模控制, 但是此类方法是基于梯度的算法, 容易陷入局部极小值, 从而产生次优的结果, 并且随着模型规模的增加, 算法的计算效率较低[10]. 另一类常用的特征选择方法是系数收缩方法, 也称为正则化方法. 文献 [11] 提出基于 L_0 范数特征选择的支持向量机模型, 提供一种有效的稀疏建模方法, 提高了模型的分类准确率和泛化性能. 文献 [12] 提出基于 L_1 范数正则化的极限学习机, 有效简化模型结构, 并且具有较高的预测精度. 此外, 学者还提出基于 L_p 范数正则化[13] 以及不同范数组合[14] 的模型特征选择方法, 同样取得了良好的效果. 因此, 借助于正则化方法的特征选择能力, 可以有效控制模型的复杂程度, 从而提高模型的可解释性[15].

基于以上分析, 本文基于 L_1 范数正则化提出一种改进的回声状态网络 (L_1-ESN). 该方法在二次损失函数的基础上, 引入 L_1 范数惩罚项, 收缩模型系数, 提高模型求解的稳定性. 同时, 通过对储备池输出特征进行选择, 剔除冗余特征和无关特征, 从而控制储备池的规模, 提高网络的泛化性能. 在模型求解过程中, 为了避免估计正则化参数, 本文采用最小角回归 (Least angle regression, LARS) 算法对模型特征进行排序, 计算完整的正则化路径, 然后根据信息准则选取最优模型, 有效降低算法的计算复杂度.

1　基于 L_1 正则化的回声状态网络

1.1　回声状态网络

回声状态网络的状态方程为[6]

$$\begin{cases} \boldsymbol{x}(t+1) = f\left(W_{\text{in}}\boldsymbol{u}(t+1) + W_x\boldsymbol{x}(t)\right) \\ y(t+1) = W_{\text{out}}^{\text{T}}\boldsymbol{x}(t+1) \end{cases} \tag{1}$$

其中, $\boldsymbol{u}(t) \in \mathbf{R}^L$、$\boldsymbol{x}(t) \in \mathbf{R}^M$ 和 $y(t) \in \mathbf{R}$ 分别表示储备池 t 时刻的输入变量、状态变量和输出变量; $W_{\text{in}} \in \mathbf{R}^{M\times L}$、$W_x \in \mathbf{R}^{M\times M}$ 和 $W_{\text{out}} \in \mathbf{R}^M$ 分别表示输入权值矩阵、内部连接权值矩阵和输出权值矩阵, 其中 W_{in} 和 W_x 全部随机产生, 并且在训练过程中保持不变, 只有储备池的输出权值 W_{out} 通过训练得到; $f(\cdot)$ 表示内部神经元的激活函数, 通常情况下取双曲正切函数.

给定训练样本集 $\{\boldsymbol{u}(t), y(t)\}_{t=1}^N$, 其中 N 表示训练数据集的规模. ESN 网络的训练过程就是确定系统输出权值 W_{out} 的过程. 在网络稳定之前的网络状态变量容易受到初始状态因素影响, 从而影响网络的最终性能, 因此常常需要舍弃初始的暂态过程. 储备池的状态矩阵 X 和对应的期望输出 $\boldsymbol{y}$ 可以表示为

$$X = [\boldsymbol{x}(T_0+1), \boldsymbol{x}(T_0+2), \cdots, \boldsymbol{x}(N)]^{\text{T}} \tag{2}$$

$$\boldsymbol{y} = [y(T_0+1), y(T_0+2), \cdots, y(N)]^{\text{T}} \tag{3}$$

其中, T_0 为舍弃的初始暂态过程长度, N 表示训练样本数, 则实际训练长度为 $\tilde{N} = N - T_0$, $X \in \mathbf{R}^{\tilde{N}\times M}$, $\boldsymbol{y} \in \mathbf{R}^{\tilde{N}}$. 因此, 回声状态网络的训练过程可以转化为求解如下的回归方程:

$$\boldsymbol{y} = XW_{\text{out}} \tag{4}$$

采用伪逆法求解式 (4) 所示的回归方程, 网络的目标函数和输出权值的表达式如下:

$$L(\hat{W}_{\text{out}}) = \|\boldsymbol{y} - XW_{\text{out}}\|_2^2 \tag{5}$$

$$\hat{W}_{\text{out}} = X^{\dagger}\boldsymbol{y} = (X^{\text{T}}X)^{-1}X^{\text{T}}\boldsymbol{y} \tag{6}$$

其中, $L(\cdot)$ 为目标函数, $\hat{W}_{\text{out}}$ 表示输出权值的估计值, $\|\cdot\|_2$ 表示 L_2 范数, $X^{\dagger}$ 表示矩阵 X 的伪逆.

采用奇异值分解方法对输出权值进行求解. 首先对储备池的状态变量矩阵 X 进行奇异值分解:

$$X = U\Sigma V^{\text{T}} \tag{7}$$

其中, 矩阵 $X \in \mathbf{R}^{\tilde{N}\times M}$, $U \in \mathbf{R}^{\tilde{N}\times P}$, $V \in \mathbf{R}^{M\times P}$, $\Sigma = \text{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_P\} \in \mathbf{R}^{P\times P}$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_P > 0$ 为矩阵 X 的非零奇异值. 则矩阵 X 的伪逆可以表示为

$$X^{\dagger} = V\Sigma^{-1}U^{\text{T}} \tag{8}$$

因此, 输出权值的伪逆解可以表示为

$$\hat{W}_{\text{out}} = V\Sigma^{-1}U^{\text{T}}\boldsymbol{y} = \sum_{i=1}^{P}\frac{1}{\sigma_i}\boldsymbol{v}_i\boldsymbol{u}_i^{\text{T}}\boldsymbol{y} \tag{9}$$

其中, $\boldsymbol{v}_i$ 和 $\boldsymbol{u}_i$ 表示矩阵 V 和 U 的列向量. 由式 (9) 可以看出, 当矩阵 X 的最小奇异值非常接近于零时, 网络的输出权值具有较大幅值, 容易产生病态解, 影响网络的稳定性. 伪逆法具有实现简单、计算效率高等优点, 是输出权值的无偏估计, 但是通常会产生较大的方差, 导致模型的训练误差很小, 而测试误差很大, 严重影响模型的泛化性能.

1.2　L_1 范数正则化模型

正则化方法可以有效解决病态解问题. 考虑在网络的目标函数中添加惩罚项, 如下式所示:

$$L(\hat{W}_{\text{out}}, \lambda) = \|\boldsymbol{y} - XW_{\text{out}}\|_2^2 + \lambda\|W_{\text{out}}\| \tag{10}$$

其中, 前一项为误差项, 后一项为惩罚项, λ 为正则化参数, 控制模型惩罚项的大小. 正则化方法可以有效平衡模型的偏差和方差, 通过引入小的偏差 (即惩罚项), 可以使模型预测的方差大幅降低, 从而提高模型的泛化性能.

L_2 范数正则化是最常用的正则化模型, 也称为岭回归, 其目标函数的惩罚项为输出权值的 L_2 范数. 由于 L_2 范数正则化是严格的凸优化问题, 输出权值的解具有解析形式:

$$\hat{W}_{\text{out}} = (X^{\text{T}}X + \lambda I)^{-1}X^{\text{T}}\boldsymbol{y} \tag{11}$$

由式 (11) 可以看出, 该模型在矩阵 $X^{\text{T}}X$ 对角元素上增加一个正常数, 避免出现奇异矩阵. 因此, L_2 范数正则化模型可以消除线性回归中的病态解问题, 改善输出权值解的性能. 但是, L_2 范数正则化模型不具有特征选择能力, 不能控制模型规模.

L_1 范数正则化模型, 也称为 LASSO (Least absolute selection and shrinkage operator)[16], 其惩罚项选择 L_1 范数, 具有 L_2 范数正则化的全部优点, 可以有效提高模型求解的数值稳定性. 此外, L_1 范数正则化具有特征选择能力, 可以产生稀疏解, 使网络无关节点的输出权值趋近于零, 从而控制网络的复杂程度, 提高模型的可解释性[17]. 与传统的子集选择方法相比, 它具有低方差的优点, 有效平衡模型预测的偏差和方差, 提高模型的泛化性能. 由于 L_1 范数正则化模型的目标函数不可微, 使得求解过程十分复杂. 正则化参数的大小控制解的稀疏性, 对模型性能影响很大, 因此需要有效确定正则化参数. 目前, 正则化参数选择方案, 主要有交叉检验、Bootstrap 方法等, 计算效率较低, 在实际应用中受到很大限制.

2 L_1-ESN 模型求解

本文基于 L_1 范数正则化提出改进的回声状态网络模型, 有效解决模型规模控制问题, 同时提高模型求解的稳定性和泛化能力. 对于模型的求解, Efron 等提出最小角回归算法, 可以有效解决 L_1 范数正则化模型的求解问题[18]. LARS 算法具有与前向选择算法相同的计算复杂度, 保证了模型求解的快速性, 在实际应用中容易实现. 下面将首先提出 L_1-ESN 模型的具体实现步骤, 然后给出 LARS 算法的转化和模型选择过程.

2.1 L_1-ESN 模型具体实现步骤

在 ESN 网络训练过程中, 考虑在目标函数中加入 L_1 范数惩罚项, 目标函数形式如式 (10) 所示, 构成 L_1-ESN 模型. 相比于传统的 ESN 模型, L_1-ESN 模型能够提高模型求解的稳定性, 同时可以有效控制模型的复杂度, 防止出现过拟合. 为了快速、有效地求解 L_1-ESN 模型, 本文提出如图 1 所示的模型求解过程.

图 1 L_1-ESN 模型实现过程

Fig. 1 Implementation steps of L_1-ESN model

L_1-ESN 模型的实现过程如图 1 所示, 其中网络的输出变量排序和模型选择过程, 等价于 L_1 范数正则化的求解. 本文采用最小角回归算法计算模型的正则化路径, 得到候选模型, 并根据贝叶斯信息准则从候选模型中选择最优模型. 具体实现步骤如下:

步骤 1. 建立 ESN 网络: 设定网络参数, 包括储备池规模、稀疏度和谱半径等; 随机初始化网络输入权值 W_{in} 和内部连接权值 W_x.

步骤 2. 状态变量采样: 选择网络的训练样本, 根据状态方程更新网络的状态变量, 构成状态变量矩阵 X.

步骤 3. 变量排序: 根据状态变量和输出变量的回归关系, 采用 LARS-LASSo 算法对状态变量进行排序, 得到完整的正则化路径, 即一组候选模型.

步骤 4. 模型选择: 根据贝叶斯信息准则, 从候选模型中选出最优模型.

2.2 最小角回归算法

LARS 算法基本思想如下: 先设所有候选变量的系数为零, 从中选择一个与响应变量相关性最大的变量, 添加到模型中, 然后沿着最小角方向依次添加最相关变量, 直至满足一定的停止准则或候选变量全部加入到预测模型中.

但是该方法没有考虑变量间的相互关系, 不能从已选变量中剔除冗余或不相关变量, 因此容易产生次优的结果. 针对以上问题, 可以修改 LARS 算法的计算流程, 从而改善解的性质. 将 LARS 算法做如下的修正:

设输入变量集合为 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_R\}$, 有效集为 $A \subset X$, 无效集为 $I = X - A$, 输出变量为 $\boldsymbol{y}$. 在前向选择过程中, 选择与当前残差 $\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}_k$ 最相关的变量 $\boldsymbol{x}_i \in I$, 步长 γ_i 计算如下:

$$\gamma_i = \min_{\boldsymbol{x}_i \in I} \left\{ \frac{c - \boldsymbol{x}_i^{\text{T}}\boldsymbol{\varepsilon}}{c - \boldsymbol{x}_i^{\text{T}}\boldsymbol{d}}, \frac{c + \boldsymbol{x}_i^{\text{T}}\boldsymbol{\varepsilon}}{c + \boldsymbol{x}_i^{\text{T}}\boldsymbol{d}} \right\} \tag{12}$$

其中, $\boldsymbol{d} = \boldsymbol{y}_{k+1} - \boldsymbol{y}_k$ 为方向向量, $c = \max_{\boldsymbol{x}_j \in A} \left| \boldsymbol{x}_j^{\text{T}}\boldsymbol{\varepsilon} \right|$, $\boldsymbol{x}_j$ 为有效集中与当前残差最相关的变量. 计算步长 γ_i 的目的是找出无效集中与当前残差最相关的变量, 将其加入到有效集中.

在后向消除过程中, 计算步长 γ_j 如下所示:

$$\gamma_j = \min_{\boldsymbol{x}_j \in A} \left(-\frac{\hat{W}_k}{\nabla \hat{W}_k} \right) \tag{13}$$

其中, $\hat{W}_k$ 表示当前的系数矩阵.

如果 $\gamma_j < \gamma_i$, 则 LARS 算法停止向前搜索, 并且从模型中移除第 j 个变量, 即

$$A^* = A - \{\boldsymbol{x}_j\} \tag{14}$$

式 (14) 表示将变量 $\boldsymbol{x}_j$ 从有效集中移除, 并添加到无效集中, 即后向消除步骤.

学者证明了 LASSO 方法与 LARS 算法之间的联系, 即经过上面的修正, LARS 算法可以得到 L_1 范数正则化模型的全部解集, 称之为 LARS-LASSO 算法[18]. 可以看出, 修正的 LARS 算法相比于传统的前向选择算法, 增加了后向消除步骤, 可以有效解决变量相关性问题, 避免产生次优的结果, 具有更理想的变量子集搜索性能. LARS-LASSO 算法计算流程如下:

步骤 1. 模型初始化: 初始化系数矩阵 $W_0 = \mathbf{0}$, 拟合向量 $\boldsymbol{y}_0 = \mathbf{0}$, 定义有效集 $A = \Phi$, 无效集 $I = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_R\}$. 找到与响应变量最相关变量, 将其从无效集 I 中移除, 添加到有效集 A.

步骤 2. 计算步长: 分别计算无效集和有效集中变量的最小步长, 记作 γ_i 和 γ_j. 其中 $\boldsymbol{x}_i \in I$, $\boldsymbol{x}_j \in A$.

步骤 3. 更新模型有效集和无效集: 如果 $\gamma_j < \gamma_i$, 则变量 $\boldsymbol{x}_j$ 从有效集 A 移除, 添加到无效集 I 中; 如果 $\gamma_j > \gamma_i$, 则变量 $\boldsymbol{x}_i$ 从无效集 I 移除, 添加到有效集 A 中.

步骤 4. 更新模型系数向量: 更新模型系数矩阵 $W_{k+1} = W_k + \gamma \nabla W_k$. 重复步骤 2 ~ 4, 直到满足某个停止准则或者全部变量添加到有效集中.

步骤 5. 输出系数序列: 根据前面计算的模型系数矩阵, 得到完整的正则化路径 $B = [W_0, \cdots, W_S]$.

2.3　基于信息准则的模型选择

根据 LARS-LASSO 算法计算得到完整的正则化路径, 构成一组候选模型. 模型选择的目的是从中选出最优的有效集和系数矩阵, 以确定最终的模型结构. 在模型选择中, 目前应用最广泛的是 Akaike 信息准则 (Akaike information criterion, AIC) 和贝叶斯信息准则 (Bayesian information criterion, BIC)[19].

AIC 准则为极大似然准则的扩展, 是评估统计模型的复杂程度和衡量统计模型优良的一种标准. 但是, AIC 准则存在一定的局限性, AIC 准则倾向于选择过多候选变量, 容易导致模型过拟合. 为了弥补 AIC 准则的不足, Schwarz 提出了 BIC 准则, 表达式如下所示:

$$\mathrm{BIC}(p) = n \log \frac{\mathrm{RSS}(p)}{n} + p \log n \tag{15}$$

其中, n 为候选变量个数, p 为有效变量个数, $\mathrm{RSS}(p) = \left\| \boldsymbol{y} - X_p \hat{W}_p \right\|_2^2$ 表示包含 p 个有效变量的模型残差平方和. 在表达式中, 第一项表示模型拟合的优良性; 第二项表示对模型复杂度的惩罚. 与 AIC 准则相比, BIC 准则加强了模型复杂度的惩罚, 倾向于选择更少的候选变量. BIC 准则是对候选模型贝叶斯后验概率的渐近逼近, 对于大样本情况, 采用 BIC 准则选出的模型更为理想, 模型可信度更高[20]. 基于以上分析, 本文采用 BIC 准则进行模型选择, 平衡模型的拟合效果和复杂度.

3　仿真实验

为了验证本文方法的有效性, 将 L_1-ESN 模型应用于 Lorenz 多元混沌时间序列、大连市气温降雨时间序列和 Gas Furnace 时间序列预测中, 并与传统的 ESN 模型[1]、基于岭回归的 ESN 模型 (Ridge echo state network, RESN)[6]、基于前向选择的 ESN 模型 (Forward selection echo state network, FS-ESN)[8] 和基于最小角回归的 ESN 模型 (Least angle regression echo state network, LAR-ESN)[8] 进行比较. 仿真实验中, 分别选取均方根误差 (Root mean square error, RMSE) 和对称平均绝对百分率误差 (Symmetric mean absolute percentage error, SMAPE) 两种性能指标, 定量评价模型的预测性能. 为了使仿真实验结果更具说服力, 本文仿真实验均通过 50 次重复实验, 取平均值.

均方根误差反映预测值相对于真实值的偏离程度, 值大于或等于零, 值越小表明预测的效果越好. 均方根误差定义为

$$\mathrm{RMSE} = \left[\frac{1}{n-1} \sum_{k=1}^{n} [\hat{y}(k) - y(k)]^2 \right]^{\frac{1}{2}} \tag{16}$$

其中, $\hat{y}(k)$ 为预测值, $y(k)$ 为真实值, n 为样本数.

对称平均绝对百分率误差衡量时间序列拟合的准确度, 主要是对时间序列趋势的评估. 对称平均绝对百分率误差定义为

$$\mathrm{SMAPE} = \frac{1}{n} \sum_{k=1}^{n} \frac{|y(k) - \hat{y}(k)|}{y(k) + \hat{y}(k)} \tag{17}$$

3.1 Lorenz 多元混沌时间序列预测

Lorenz 系统表达式如下所示:

$$\begin{cases} \dot{x} = a(-x+y) \\ \dot{y} = bx - y - xz \\ \dot{z} = xy - cz \end{cases} \tag{18}$$

其中, 当初始值 $[x(0), y(0), z(0)] = [12, 2, 9]$, 参数 $[a, b, c] = [10, 28, 8/3]$ 时, Lorenz 时间序列具有混沌特性. 利用四阶龙格–库塔法求解式 (18), 步长为 0.02, 得到 2 501 组 Lorenz 多元时间序列数据. 本文将 $[x(t), y(t)]^{\mathrm{T}}$ 作为预测模型的输入变量, $x(t+1)$ 作为输出变量. 首先对输入变量进行相空间重构, 应用关联维数法计算得到重构参数为 $\tau_x = 2, m_x = 8, \tau_y = 1, m_y = 8$, 因此模型的输入变量 $\boldsymbol{u}(t)$ 为

$$\boldsymbol{u}(t) = [x(t), \cdots, x(t-7\tau_x),\ y(t), \cdots, y(t-7\tau_y)]^{\mathrm{T}} \tag{19}$$

经过相空间重构, 得到 2 487 组重构样本 $\boldsymbol{u}(t)$, 被评估的模型采用前 2 000 组样本进行训练, 余下的 487 组样本进行测试, 前 100 组样本用于消除储备池初始暂态的影响. 为了验证本文模型的特征选择性能, 分别采用不同规模的储备池对 Lorenz 多元混沌时间序列进行预测. 部分预测结果如表 1 所示, 不同预测模型的均方根误差曲线如图 2 所示.

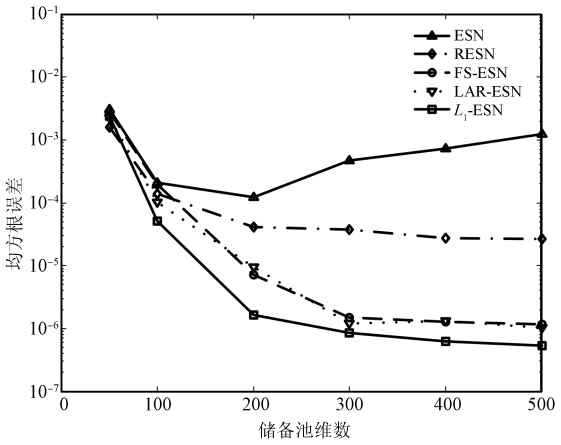


图 2 5 种模型 Lorenz 时间序列预测误差比较

Fig. 2 Prediction error of five models for Lorenz time series

由表 1 的预测结果可以看出, 本文提出的 L_1-ESN 模型与其他模型相比, 预测精度有较大优势. 模型规模指网络输出权值中非零权值的个数, 反映模型的复杂程度. ESN 和 RESN 模型的学习算法不具有特征选择能力, 因此不能控制模型规模. FS-ESN、LAR-ESN 和 L_1-ESN 模型借助于学习算法的特征选择过程, 可以有效控制模型复杂程度, 从而提高模型的泛化性能. 由于 BIC 准则倾向于选择复杂度较低的模型, 可以有效平衡模型的拟合精度和复杂度, 因此模型复杂度有明显降低. 其中, 本文提出的 L_1-ESN 模型在模型规模控制方面取得了最好的效果, 网络输出权值中非零权值个数最少.

表 1 不同模型 Lorenz 时间序列预测结果比较

Table 1 Comparison of different models for Lorenz time series prediction

储备池维数	预测模型	RMSE	SMAPE	模型规模
100	ESN	2.07E−4	7.13E−5	100
	RESN	1.39E−4	5.62E−5	100
	FS-ESN	1.96E−4	8.76E−5	100
	LAR-ESN	1.03E−4	5.61E−5	100
	L_1-ESN	5.17E−5	2.12E−5	97
300	ESN	4.71E−4	8.68E−5	300
	RESN	3.76E−5	9.60E−6	300
	FS-ESN	1.48E−6	6.16E−7	269
	LAR-ESN	1.30E−6	4.09E−7	282
	L_1-ESN	8.47E−7	2.94E−7	221
500	ESN	1.24E−3	4.57E−4	500
	RESN	2.66E−5	9.03E−6	500
	FS-ESN	1.17E−6	4.73E−7	303
	LAR-ESN	1.04E−6	4.14E−7	306
	L_1-ESN	5.31E−7	1.76E−7	261

在图 2 中, 横坐标表示不同的储备池维数, 纵坐标表示 Lorenz 多元时间序列预测均方根误差. 可以看出, 本文提出的 L_1-ESN 模型获得了最小的误差. 当储备池维数大于 300 时, ESN 模型开始出现过拟合现象, 预测误差逐渐增加. RESN 模型通过引入 L_2 范数惩罚项, 模型具有良好的稳定性, 但是模型预测精度没有明显提高. 本文提出的 L_1-ESN 模型通过引入 L_1 范数惩罚项, 与传统 ESN 模型相比, 可以提高模型求解的稳定性, 同时借助于模型的特征选择能力, 选择出对模型影响最大的特征子集, 提高了模型的预测性能, 可以有效避免过拟合.

3.2 大连市气温降雨时间序列预测

选择实际观测的大连市气温降雨时间序列, 数据包括 1951 年 1 月到 2010 年 7 月的大连市月平均气温和降雨量, 共计 715 组数据. 由于原始数据包含大量的噪声, 不利于构建预测模型, 因此首先采

用奇异谱分析方法对原始数据进行去噪处理[21], 得到序列的主要演变特征. 然后, 对去噪后的数据进行相空间重构, 由于气象数据具有明显的周期性, 选取嵌入维数为 12, 延迟时间为 1, 得到模型的输入变量为

$$\boldsymbol{u}(t) = [x_1(t), \cdots, x_1(t-11), x_2(t), \cdots, x_2(t-11)]^{\mathrm{T}} \tag{20}$$

其中, $x_1(t)$ 和 $x_2(t)$ 分别为大连市月平均气温和降雨量, 输入变量 $\boldsymbol{u}(t)$ 共 24 维. 根据输入变量分别预测 $x_1(t+1)$ 和 $x_2(t+1)$. 预测过程中, 选取前 500 组数据进行训练, 余下的 203 组数据进行测试, 储备池维数为 200.

大连市月平均气温和降雨量一步预测结果如图 3 和图 4 所示, 可以看出本文提出的 L_1-ESN 模型对时间序列的拟合效果良好, 能够真实地跟踪气温和降雨量时间序列的走势. 不同模型预测误差比较如表 2 和表 3 所示. 由表 2 和表 3 的预测结果可以看出, 本文提出的 L_1-ESN 模型在实际观测数据预测中同样具有良好的预测效果. ESN 和 RESN 模型预测结果相近, 二者容易受到噪声和异常点的影响, 鲁棒性较差. 通过引入特征选择过程, 可以剔除 ESN 网络的冗余特征和无关特征, 降低噪声数据对预测模型的影响, 从而提高网络的泛化性能, 因此本文所提模型具有较高预测精度. 此外, 本文模型具有良好的稳定性, 可以有效避免过拟合现象.

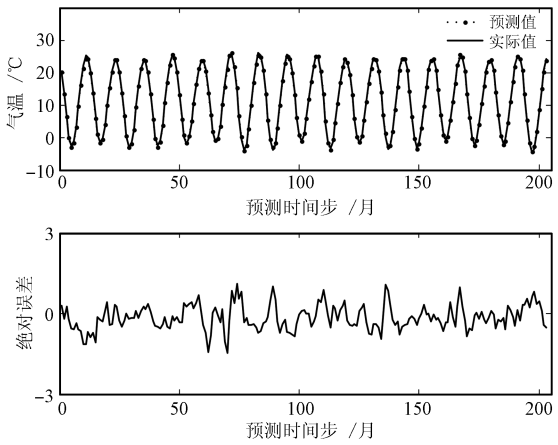


图 3　月平均气温预测结果

Fig. 3　Prediction results for average monthly temperature

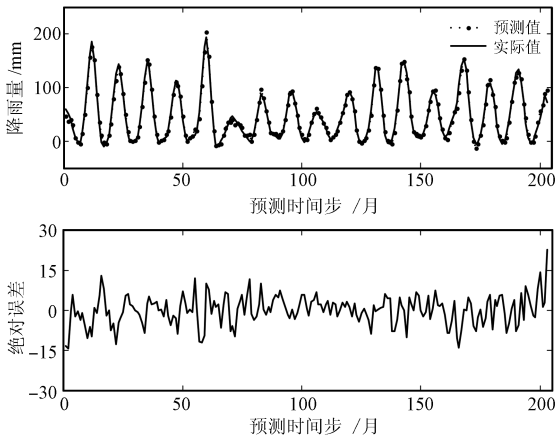


图 4　月平均降雨量预测结果

Fig. 4　Prediction results for average monthly rainfall

表 2　月平均气温预测误差比较

Table 2　Prediction error comparison for average monthly temperature

预测模型	RMSE	SMAPE
ESN	0.6785	0.2351
RESN	0.5571	0.2516
FS-ESN	0.5210	0.2344
LAR-ESN	0.5196	0.1059
L_1-ESN	0.4302	0.1035

表 3　月平均降雨量预测误差比较

Table 3　Prediction error comparison for average monthly rainfall

预测模型	RMSE	SMAPE
ESN	8.3271	0.4147
RESN	8.2331	0.3379
FS-ESN	7.6252	0.2723
LAR-ESN	7.2972	0.3677
L_1-ESN	6.3084	0.2716

3.3　Gas Furnace 时间序列预测

Gas Furnace[22] 是一组常用的工业时间序列数据, 输入变量为气体速率 $x(t)$, 输出变量为 CO_2 浓度 $y(t)$, 共计 296 组样本. 首先, 进行相空间重构, 选取延迟时间为 1, 嵌入维数分别为 6 和 4, 得到模型的输入变量为

$$\boldsymbol{u}(t) = [x(t), \cdots, x(t-5), y(t), \cdots, y(t-3)]^{\mathrm{T}} \tag{21}$$

经过相空间重构, 得到 291 组重构样本, 采用前 200 组样本进行训练, 余下的 91 组样本进行测试. 根据输入变量 $\boldsymbol{u}(t)$, 预测 CO_2 浓度 $y(t+1)$. 不同模型的预测结果如表 4 所示.

表 4 Gas Furnace 时间序列预测误差比较

Table 4 Prediction error comparison for Gas Furnace time series

预测模型	RMSE	方差
ESN	1.1600	0.0603
RESN	1.1162	0.0437
FS-ESN	0.8285	0.0211
LAR-ESN	0.7362	0.0111
L_1-ESN	0.7064	0.0059

从表 4 可以看出, L_1-ESN 取得了最小的预测误差. 在实验过程中, 由于 Gas Furnace 数据含有大量噪声干扰, 传统 ESN 模型和 RESN 模型预测误差较大, 并且具有较大方差. 与其他方法相比, 本文所提方法预测结果的方差更小, 具有良好的稳定性. 实验证明, 本文方法在实际观测的工业时间序列预测中, 同样具有较好的预测结果.

4 结论

本文提出一种基于 L_1 范数正则化的改进回声状态网络. 本文方法有效解决了病态解问题, 提高了模型求解的数值稳定性. 借助于 L_1 范数正则化的特征选择能力, 对储备池的输出特征进行选择, 可以控制网络的复杂程度, 提高模型的泛化性能. 本文分别对三组时间序列进行仿真实验, 证明了本文所提模型的有效性与实用性. 与其他预测模型对比, 本文方法具有良好的稳定性, 取得了更好的预测结果.

References

1 Jaeger H, Hass H. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science*, 2004, **304**(5667): 78−80

2 Qiao Jun-Fei, Bo Ying-Chun, Han Guang. Application of ESN-based multi indices dual heuristic dynamic programming on wastewater treatment process. *Acta Automatica Sinica*, 2013, **39**(7): 1146−1151
(乔俊飞, 薄迎春, 韩广. 基于 ESN 的多指标 DHP 控制策略在污水处理过程中的应用. 自动化学报, 2013, **39**(7): 1146−1151)

3 Ongenae F, Van Looy S, Verstraeten D, Verplancke T, Benoit D, De Turck F, Dhaene T, Schrauwen B, Decruyenaere J. Time series classification for the prediction of dialysis in critically ill patients using echo state networks. *Engineering Applications of Artificial Intelligence*, 2013, **26**(3): 984−996

4 Li G Q, Niu P F, Zhang W P, Zhang Y. Control of discrete chaotic systems based on echo state network modeling with an adaptive noise canceler. *Knowledge-Based Systems*, 2012, **35**: 35−40

5 Peng Yu, Wang Jian-Min, Peng Xi-Yuan. Researches on time series prediction with echo state networks. *Acta Electronica Sinica*, 2010, **38**(2A): 148−154
(彭宇, 王建民, 彭喜元. 基于回声状态网络的时间序列预测方法研究. 电子学报, 2010, **38**(2A): 148−154)

6 Lukosevicius M, Jaeger H. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 2009, **3**(3): 127−149

7 Rong H J, Ong Y S, Tan A H, Zhu Z. A fast pruned-extreme learning machine for classification problem. *Neurocomputing*, 2008, **72**(1−3): 359−366

8 Dutoit X, Schrauwen B, Van Campenhout J, Stroobandt D, Van Brussel H, Nuttin M. Pruning and regularization in reservoir computing. *Neurocomputing*, 2009, **72**(7−9): 1534−1546

9 Miche Y, Sorjamaa A, Bas P, Simula O, Jutten C, Lendasse A. OP-ELM: optimally pruned extreme learning machine. *IEEE Transactions on Neural Networks*, 2010, **21**(1): 158−162

10 Kump P, Bai E W, Chan K S, Eichinger B, Li K. Variable selection via RIVAL (removing irrelevant variables amidst LASSO iterations) and its application to nuclear material detection. *Automatica*, 2012, **48**(9): 2107−2115

11 Liu Qiao, Qin Zhi-Guang, Chen Wei, Zhang Feng-Li. Zero-norm penalized feature selection support vector machine. *Acta Automatica Sinica*, 2011, **37**(2): 252−256
(刘峤, 秦志光, 陈伟, 张凤荔. 基于零范数特征选择的支持向量机模型. 自动化学报, 2011, **37**(2): 252−256)

12 Han Min, Li De-Cai. An norm 1 regularization term ELM algorithm based on surrogate function and Bayesian framework. *Acta Automatica Sinica*, 2011, **37**(11): 1344−1350
(韩敏, 李德才. 基于替代函数及贝叶斯框架的 1 范数 ELM 算法. 自动化学报, 2011, **37**(11): 1344−1350)

13 Liu Jian-Wei, Li Shuang-Cheng, Luo Xiong-Lin. Classification algorithm of support vector machine via p-norm regularization. *Acta Automatica Sinica*, 2012, **38**(1): 76−87
(刘建伟, 李双成, 罗雄麟. p范数正则化支持向量机分类算法. 自动化学报, 2012, **38**(1): 76−87)

14 Miche Y, Van Heeswijk M, Bas P, Simula O, Lendasse A. TROP-ELM: a double-regularized ELM using LARS and Tikhonov regularization. *Neurocomputing*, 2011, **74**(16): 2413−2421

15 Friedman J H. Fast sparse regression and classification. *International Journal of Forecasting*, 2012, **28**(3): 722−738

16 Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, **58**(1): 267−288

17 Peng Yi-Gang, Suo Jin-Li, Dai Qiong-Hai, Xu Wen-Li. From compressed sensing to low-rank matrix recovery: theory and applications. *Acta Automatica Sinica*, 2013, **39**(7): 981−994
(彭义刚, 索津莉, 戴琼海, 徐文立. 从压缩传感到低秩矩阵恢复: 理论与应用. 自动化学报, 2013, **39**(7): 981−994)

18 Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *The Annals of Statistics*, 2004, **32**(2): 407−499

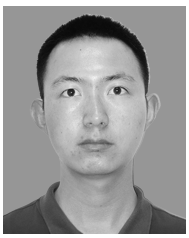
19 Stoica P, Selen Y. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 2004, **21**(4): 36−47

20 Watanabe S. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 2013, **14**(1): 867−897

21 Wu C L, Chau K W. Prediction of rainfall time series using modular soft computing methods. *Engineering Applications of Artificial Intelligence*, 2013, **26**(3): 997−1007

22 Box G E P, Jenkins G M, Reinsel G C. *Time Series Analysis: Forecasting and Control*. New Jersey, USA: John Wiley & Sons, 2008. 677−678

韩 敏 大连理工大学电子信息与电气工程学部教授. 主要研究方向为神经网络, 模式识别, 混沌时间序列预测. 本文通信作者.
E-mail: minhan@dlut.edu.cn
(**HAN Min** Professor at the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. Her research interest covers neural networks, pattern recognition, and chaotic time series prediction. Corresponding author of this paper.)



任伟杰 大连理工大学电子信息与电气工程学部博士研究生. 主要研究方向为变量选择, 多元时间序列预测.
E-mail: renweijie@mail.dlut.edu.cn
(**REN Wei-Jie** Ph. D. candidate at the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. His research interest covers variable selection and multivariate time series prediction.)

许美玲 大连理工大学电子信息与电气工程学部博士研究生. 研究方向为神经网络和多元时间序列预测.
E-mail: xuml@mail.dlut.edu.cn
(**XU Mei-Ling** Ph. D. candidate at the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. Her research interest covers neural networks and multivariate time series prediction.)