

基于改进黑洞算法优化 ESN 的网络流量短期预测

韩莹^{1,2}, 井元伟¹, 金建宇³, 李琨²

(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110819; 2. 渤海大学 工学院, 辽宁 锦州 121013;

3. 东北大学 国防教育学院, 辽宁 沈阳 110819)

摘 要: 网络流量数据序列具有混沌特性. 相空间重构后, 采用一种改进黑洞算法优化回声状态网络的非线性模型对网络流量进行预测. 改进黑洞算法是在现有工作的基础上提出一种新的新解生成机制, 可以提高算法的收敛速度和精度; 相比于遗传算法、和声搜索算法等其他优化算法, 所提出的改进黑洞算法不依赖自身相关参数的准确设定; 将其应用于回声状态网络 4 个重要参数的优化选取, 使得预测模型具有较好的预测稳定性. 通过 Mackey-Glass 混沌时间序列和网络流量公共数据集的仿真实验, 结果表明所提出的方法具有较好的预测性能.

关 键 词: 网络流量; 混沌时间序列; 回声状态网络; 黑洞算法; 预测

中图分类号: TP 393

文献标志码: A

文章编号: 1005-3026(2018)03-0311-05

Network Traffic Short-Term Prediction Based on Echo State Network Optimized by Improved Black Hole Algorithm

HAN Ying^{1,2}, JING Yuan-wei¹, JIN Jian-yu³, LI Kun²

(1. School of Information Science & Engineering, Northeastern University, Shenyang 110819, China; 2. College of Engineering, Bohai University, Jinzhou 121013, China; 3. School of National Defense Education, Northeastern University, Shenyang 110819, China. Corresponding author: JING Yuan-wei, E-mail: jingyuanwei@ise.neu.edu.cn)

Abstract: The network traffic data series has chaos characteristics. After phase space reconstruction, a nonlinear prediction model based on echo state network (ESN) optimized by improved black hole (BH) was used to predict network traffic. The improved BH algorithm is a new mechanism for new-solution generation based on current works, which can increase the algorithm's convergence speed and precision. Compared with other optimization algorithms, such as genetic algorithm (GA), harmony search (HS) algorithm, etc, the proposed improved BH algorithm is not affected by the accuracy of the setting for some parameters of itself. It is used to optimally select four key parameters of the ESN model, which has better prediction stability. Simulation experiments of Mackey-Glass chaos time series and public network traffic data set show that the proposed method has better prediction ability.

Key words: network traffic; chaos time series; echo state network (ESN); black hole algorithm; prediction

网络流量是评价网络负载和运行状态的一个较为重要的参数, 通过对其连续监测以实现准确预测, 是掌握网络运行状况以及实施有效管理和控制的重要手段. 网络流量是基于时间的数据, 具有时间序列数据的特点, 有学者已经证明其具有

混沌特性^[1], 可以对其进行预测. 对于时间序列数据的预测方法, 主要有线性方法和非线性方法. 其中, 线性方法主要有: 自回归 (auto regression, AR)^[2]、自回归滑动平均 (auto regressive moving average, ARMA)^[3]、自回归综合滑动平均 (auto

regressive integrated moving average, ARIMA)^[4]等;非线性方法主要有:支持向量机(support vector machine, SVM)^[5]、最小二乘支持向量机(least square support vector machine, LSSVM)^[6]、人工神经网络(artificial neural network, ANN)^[7]、灰色模型(grey model, GM)^[8]等. 随着网络规模和复杂程度的不断提高,非线性分析可以很好地处理复杂数据的非平稳、非线性等问题,因此,非线性方法在网络流量预测中发挥着越来越重要的作用.

回声状态网络(echo state network, ESN)是一种新型的递归神经网络,具有算法简单、收敛速度快等特点^[9]. 相对于 SVM, LSSVM, ANN 和 GM 等方法,ESN 有其独特的优势,虽然 SVM 和 LSSVM 方法需要的样本较少,但是其预测精度受限于核函数及其参数的准确选取,且易陷入局部极值;ANN 方法虽然能够以任意精度逼近函数,但是其模型结构复杂,较难精确给定;GM 方法较难处理网络流量数据存在剧烈变化的情况;ESN 方法内部设置一个动态储备池(dynamic reservoir, DR),由大量神经元构成,蕴含非线性系统的动态特性,具有短暂记忆的功能,ESN 网络的输入层到储备池的连接权值以及储备池内部的连接权值在初始化时随机生成,且在整个训练过程中保持不变,只有输出权值是需要求解的部分,使其与 ANN 相比收敛速度快,且没有陷入局部极值的问题,有学者已经将 ESN 方法应用于网络流量的预测中^[1,10].

ESN 模型的预测精度主要由 DR 的 4 个参数决定,分别为:权值谱半径(SR)、储备池大小(N)、输入尺度因子(IS)和储备池神经元连接程度(SD),这几个参数的取值对 ESN 的预测精度有重要的影响^[1],因此,在模型训练过程中实现对 4 个参数的优化选择是非常重要的. 黑洞(black hole, BH)算法是 2013 年由 Hatamlou 提出的一种启发式优化算法^[11],原理简单,易于实现. 与遗传算法(genetic algorithm, GA)、粒子群(particle swarm optimization, PSO)、模拟退火(simulated annealing, SA)、蚁群(ant colony optimization, ACO)以及和声搜索(harmony search, HS)等算法相比, BH 算法的最大优点是受优化算法自身参数设置的影响较小^[6],因为这些参数的设定一般依赖人工经验,如果设置不合理,会影响算法的计算精度. 通过仿真实验已经证明, BH 算法的计算精度优于 GA, PSO 等算法^[6,11],因此,该算法已受到很多学者的关

注^[6,12-13]. 但是, BH 算法也有它自身的不足,就是算法的运行时间较长,造成收敛速度较慢;对此,在对经典算法中新解的生成机制会影响收敛速度进行分析后,提出一种改进的黑洞算法,采用一种新的新解生成机制,在提高算法收敛速度的同时,还能够改善其收敛精度.

1 混沌时间序列

原始数据集表示为 $\{x_i, i = 1, 2, \dots, Z\}$, 相空间重构后,得到如下混沌时间序列: $X_t = [x_t, x_{t+\tau}, \dots, x_{t+(m-1)\tau}]$, $t = 1, 2, \dots, Z - (m - 1)\tau$. 其中: m 和 τ 分别为嵌入维数和延迟时间, m 和 τ 的取值可以根据 C - 法^[14]进行计算. 混沌时间序列的输出为 $Y_t = x_{t+1+(m-1)\tau}$, 将 X_t 和 Y_t 分别作为预测模型的输入和输出进行建模. 对于时间序列预测模型的步长,本文采用单步预测方式,即每预测出一个 Y_t 后,使其自动加入到新的输入 X_{t+1} 的最后一维.

2 回声状态网络(ESN)

设定系统有 M 个输入, N 个内部状态, K 个输出. 那么,它们在时刻 t 的值分别表示为

$$\left. \begin{aligned} \mathbf{u}(t) &= [u_1(t), u_2(t), \dots, u_M(t)]^T, \\ \mathbf{x}(t) &= [x_1(t), x_2(t), \dots, x_N(t)]^T, \\ \mathbf{y}(t) &= [y_1(t), y_2(t), \dots, y_K(t)]^T. \end{aligned} \right\} \quad (1)$$

其中, $\mathbf{u}(t)$, $\mathbf{x}(t)$ 和 $\mathbf{y}(t)$ 分别为输入、内部状态和输出在时刻 t 的值.

ESN 的学习步骤如下:

$$\left. \begin{aligned} \mathbf{x}(t+1) &= f(\mathbf{W}_{in} \cdot \mathbf{u}(t+1) + \mathbf{W} \cdot \mathbf{x}(t) + \mathbf{W}_{back} \cdot \mathbf{y}(t)), \\ \mathbf{y}(t+1) &= f_{out}(\mathbf{W}_{out} \cdot (\mathbf{u}(t+1), \mathbf{x}(t+1))). \end{aligned} \right\} \quad (2)$$

其中: $\mathbf{x}(t+1)$ 和 $\mathbf{y}(t+1)$ 分别为内部状态单元和输出单元在时刻 $t+1$ 的值; $f(\cdot)$ 和 $f_{out}(\cdot)$ 分别为 DR 内部神经元和输出神经元的激活函数; \mathbf{W}_{in} 为输入层到储备池的连接权值矩阵, $N \times M$ 阶; \mathbf{W} 为储备池内部单元的连接权值矩阵, $N \times N$ 阶; \mathbf{W}_{back} 为输出层反馈到储备池内部单元的连接权值矩阵, $N \times K$ 阶; \mathbf{W}_{out} 为储备池内部单元到输出层的连接权值矩阵, $K \times (M + N + K)$ 阶.

根据式(2),各单元连接权值的计算决定了 ESN 的输出结果,由于 \mathbf{W}_{in} , \mathbf{W} 和 \mathbf{W}_{back} 在初始化时随机生成,且在整个训练过程中保持不变,因此只有 \mathbf{W}_{out} 是需要求解的部分,可以通过 DR 内部

状态单元和模型期望输出的线性回归最小平均误差得到 \mathbf{W}_{out} , 有

$$\mathbf{y}(t) \approx \tilde{\mathbf{y}}(t) = \sum_{i=1}^K \mathbf{w}_{\text{out}}^i \cdot x_i(t). \tag{3}$$

其中: $\mathbf{y}(t)$ 为模型期望输出; $\tilde{\mathbf{y}}(t)$ 为模型实际输出; $\mathbf{w}_{\text{out}}^i$ 是计算权值, 为 \mathbf{W}_{out} 的元素.

在模型训练过程中, 希望得到的 $\mathbf{w}_{\text{out}}^i$ 能够使模型的均方误差最小, 即有如下的优化求解问题:

$$\min \frac{1}{P-m+1} \sum_{t=m}^P (\mathbf{y}(t) - \sum_{i=1}^K \mathbf{w}_{\text{out}}^i \cdot x_i(t))^2. \tag{4}$$

可以归纳为如下描述方式:

$$\mathbf{W}_{\text{out}} = (\mathbf{B}^{-1} \times \mathbf{L})^T. \tag{5}$$

其中: \mathbf{B} 为 $x_1(t), x_2(t), \dots, x_N(t)$ 构成的矩阵, $(P-m+1) \times N$ 阶; \mathbf{L} 为 $y_1(t), y_2(t), \dots, y_K(t)$ 构成的列矩阵, $(P-m+1) \times 1$ 阶.

3 改进的黑洞算法

3.1 黑洞 (BH) 算法

下面简要介绍 BH 算法的基本原理^[6]:

步骤 1 进行初始化, 由星星表示空间中的候选解.

步骤 2 将每个星星所代表的候选解代入适应度函数, 最优适应度函数值所对应的星星认定为黑洞.

步骤 3 所有星星向黑洞移动, 位置变化如下:

$$p_i(k+1) = p_i(k) + \text{rand}() \cdot (p_{\text{BH}} - p_i(k)). \tag{6}$$

其中: $i = 1, 2, \dots, n$, n 为候选解的数量; $p_i(k)$ 和 $p_i(k+1)$ 分别为第 i 个星星在第 k 和 $k+1$ 次迭代时的位置; p_{BH} 为黑洞的位置.

步骤 4 重新计算每个星星的适应度函数值, 如果星星比黑洞具有更优的值, 互换它们的位置.

步骤 5 如果星星移动到黑洞的半径之内, 会被其吸收, 黑洞的半径定义如下:

$$R_{\text{BH}} = \frac{f_{\text{BH}}}{\sum_{i=1}^n f_i}. \tag{7}$$

其中: f_{BH} 为黑洞的适应度函数值; $\sum_{i=1}^n f_i$ 为所有星星的适应度函数值之和.

某个星星被黑洞吸收而消失, 为了保证候选解数量的一致, 会在解空间内的随机位置上生成一个新的星星.

步骤 6 如果算法满足终止条件, 输出最优

解; 如果不满足, 返回到步骤 3 重新迭代.

3.2 BH 算法的改进

本文在文献[6]的基础上, 对 BH 算法中星星被黑洞吸收而消失后产生新解的机制进行进一步的研究, 以提高算法的优化性能. 根据 BH 算法中步骤 5, 迭代前期, 星星的移动方向较为固定, 算法的收敛速度较快; 而在后期, 当前最优解 (黑洞) 附近可能会存在更优解, 但搜索的吸收-生成机制降低了算法的局部寻优能力, 使得算法收敛变慢. 因此, 应该将生成新解的范围限定在黑洞附近, 从而增加得到更优解的可能. 对此, 本文对 BH 算法进行改进, 提出一种新解生成的新机制.

首先, 假设第 i 个星星被黑洞吸收而消失, 根据式 (8) 和 (9) 确定生成新解的范围.

$$p_i^{\text{new_inf}} = p_{\text{BH}} - \text{rand}() \cdot (p_{\text{BH}} - p_i^{\text{min}}), \tag{8}$$

$$p_i^{\text{new_sup}} = p_{\text{BH}} + \text{rand}() \cdot (p_{\text{BH}} - p_i^{\text{max}}). \tag{9}$$

其中: $p_i^{\text{new_inf}}$ 和 $p_i^{\text{new_sup}}$ 分别为生成新解范围的下界和上界; p_i^{min} 和 p_i^{max} 分别为第 i 个星星在当前代前具有最小适应度值和最大适应度值的位置.

然后, 在范围 $[p_i^{\text{new_inf}}, p_i^{\text{new_sup}}]$ 内取任意的一个值作为第 i 个星星被黑洞吸收而消失后生成的新解的位置 p_i^{new} . 可以看到, 所提出的新解生成的机制通过将生成新解的位置限定在最优解 (黑洞) 周围, 既可以使算法避免陷入局部最优解, 又能使搜索过程以较高的效率在最优解附近寻找更优解.

于是, ESN 的 4 个参数 SR, N , IS 和 SD 就可以采用本文所提出的改进 BH 算法进行优化选择.

4 基于改进 BH 算法优化 ESN 的时间序列预测

基于改进 BH 算法优化 ESN 的时间序列预测模型的主要计算步骤如下.

步骤 1 参数初始化, 设定待优化参数的取值范围^[1], 其中: $\text{SR} \in [0.1, 1)$, $N \in [10, 150)$, $\text{IS} \in [0.01, 1)$, $\text{SD} \in [0.01, 1)$.

步骤 2 将原始数据集中的数据归一化到区间 $[0, 1]$, 对其进行相空间重构, 确定模型的输入和输出, 由 ESN 方法进行训练和建模.

步骤 3 调用改进的 BH 算法进行参数的优化选择, 设置候选解的种群数量并给定初值, 将其定位为星星.

步骤 4 将每个星星所代表的参数值代入预测模型, 计算适应度函数值, 适应度函数定义

如下：

$$\text{fitness} = \sqrt{\frac{1}{Z} \sum_{i=1}^Z (y_i - \hat{y}_i)^2} \quad (10)$$

其中: Z 为时间序列的数据样本数; y_i 为实际值; \hat{y}_i 为模型预测值, 将具有最小 fitness 值的星星看作黑洞.

步骤 5 所有星星向黑洞移动, 根据式(6)进行位置的更新.

步骤 6 重新计算每个星星的 fitness 值, 如果某个星星的 fitness 值比黑洞的更小, 那么就互换两者的位置.

步骤 7 根据式(7), 计算每个星星与黑洞的距离, 如果小于 R_{BH} , 该星星在空间内消失; 同时, 根据式(8)和式(9), 由第 3.2 节所提出的新解生成的新机制在解空间内生成一个新解.

步骤 8 如果满足终止条件, 算法终止, 输出黑洞所对应的最优参数, 否则返回步骤 5 重新迭代.

步骤 9 将所得到的最优参数代入 ESN 模型中进行预测, 并对预测结果进行反归一化处理.

5 仿真分析

采用 Mackey - Glass 混沌时间序列进行仿真,

$$\text{d}x(t)/\text{d}t = [0.2x(t - \eta)]/\eta - 0.1x(t) \quad (11)$$

当 $\eta > 17$ 时, 式(11)具有混沌行为, 给定 $\eta = 30$ 和初始值 $x(t)|_{t=0} = 0.9$, 生成 600 组数据. 随机选取 400 组数据进行训练, 其余的 200 组数据进行测试.

设定混沌时间序列的参数 $m = 5$ 和 $\tau = 1$, 分别采用黑洞算法优化 ESN(BH - ESN)、改进的黑洞算法优化 ESN(IBH - ESN)和本文方法进行预测, 进化代数为 100. 图 1 为 3 种方法的实际值与预测值的对比曲线, 表 1 为两种方法的均方根误差 RMSE(root mean square error)和平均绝对误差 MAE(mean absolute error)评价指标的对比结果.

表 1 三种方法的关键性能指标对比

Table 1 Comparison of key performance indicators among the three methods

方法	RMSE	MAE
BH - ESN	0.015 8	0.011 2
IBH - ESN	0.007 9	0.005 9
本文方法	0.002 8	0.002 0

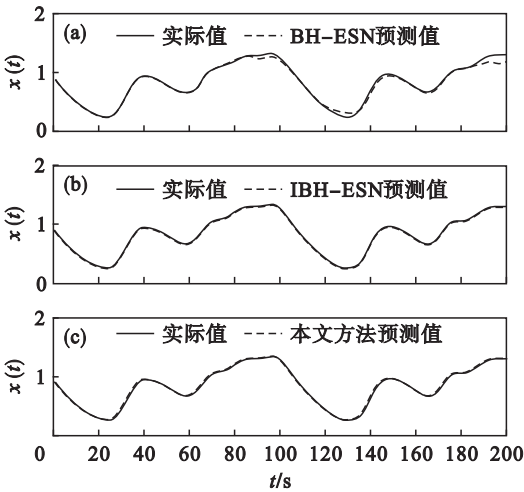


图 1 三种方法的实际值与预测值的对比
Fig. 1 Comparison of the three methods with true values and predictive values

6 实例验证

采用网络流量公共数据集 BC - pAug89 进行实例验证. 文献[1]通过对不同采样时间间隔下网络流量时间序列进行混沌特性分析和相空间重构, 已得到其相空间参数. 本文采用 T_100 ms 数据集进行验证, 采样数据集长度取 1 000, 前 400 组数据作为训练样本, 第 401 ~ 500 组数据作为测试样本, 数据集的样本序列分别如图 2 所示.

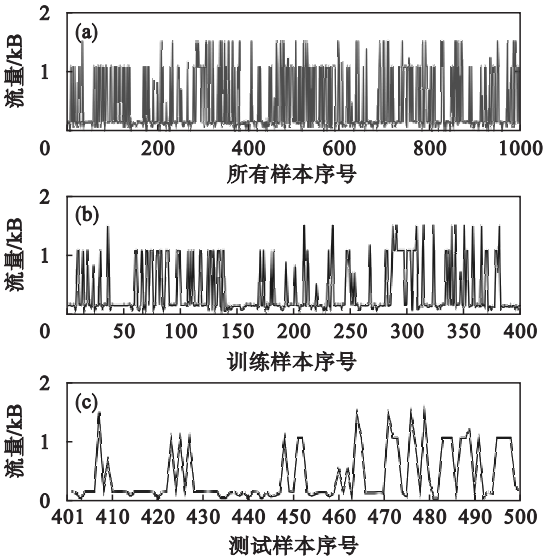


图 2 T_100 ms 数据集网络流量序列
Fig. 2 Network traffic data sequence of T_100 ms data set

将本文所提出的改进 BH - ESN 预测模型分别与 HS - ESN, GA - LSSVM, HS - LSSVM 和 Elman 神经网络预测模型进行比较. 其中, ESN 参数优化的取值范围为: $\text{SR} \in [0.1, 1)$, $N \in [10,$

150), $IS \in [0.01, 1)$ 和 $SD \in [0.01, 1)$; GA - LSSVM 中参数设置如下: 编码串长取 10, 交叉概率 $p_c \in [0.4, 0.8]$, 变异概率 $p_m \in [0.001, 0.005]$; HS - ESN 和 HS - LSSVM 中的参数设置如下: 和声数 $M \in [10, 40]$, 保留概率 $HMCR \in [0.5, 0.9]$, 记忆扰动概率 $PAR \in [0.05, 0.2]$, 带宽范围为 $[0.001, 1]$; Elman 神经网络预测模型中设置输入层 $\in [100, 150]$, 中间层 $\in [40, 80]$, 输出层 $\in [30, 50]$. 设定算法迭代次数为 100 次, 每个模型重复执行 20 次取平均(平均值 \pm 标准偏差), 每次执行过程中, 各算法的相关参数在取值范围内随机取值. 由 RMSE 和 MAE 作为评价指标. T_100ms 数据集在不同模型下的预测指标如表 2 所示. 本文所提方法预测精度不依赖参数设置, 预测结果较为稳定, 在网络流量短期预测上具有较好的效果.

表 2 T_100 ms 数据集预测指标			
Table 2 Prediction index of the T_100 ms data set			
方法	RMSE (平均值 \pm 标准偏差)	MAE (平均值 \pm 标准偏差)	
GA - LSSVM	21.918 3 \pm 31.411 9	19.533 2 \pm 26.233 0	
HS - LSSVM	16.686 7 \pm 34.123 7	14.357 4 \pm 28.635 0	
Elman 神经网络	20.062 6 \pm 34.030 7	17.559 6 \pm 28.484 2	
HS - ESN	14.259 9 \pm 34.775 1	12.001 4 \pm 29.246 2	
本文方法	4.148 2 \pm 4.067 4	3.842 6 \pm 3.619 6	

7 结 论

本文提出一种基于混沌时间序列和回声状态网络模型的网络流量短期预测方法. 首先对网络流量数据集进行相空间重构; 然后针对影响模型预测性能的四个重要参数: 权值谱半径、储备池大小、输入尺度因子和储备池神经元连接程度, 采用改进的黑洞算法进行优化选取. 所提出的改进黑洞算法受主观设定参数的影响较小, 使得预测模型不依赖于相关参数的准确设定. 根据仿真实验结果, 所提出的方法相比于其他方法, 模型的预测结果较为稳定, 且具有更好的预测性能.

参考文献:

[1] 田中大, 李树江, 王艳红, 等. 基于混沌理论与改进回声状态网络的网络流量多步预测[J]. 通信学报, 2016, 37(3): 55 - 70.
(Tian Zhong-da, Li Shu-jiang, Wang Yan-hong, et al. Network traffic multi-step prediction based on chaos theory and improved echo state network [J]. Journal on Communications, 2016, 37(3): 55 - 70.)

[2] Chisci L, Mavino A, Perferi G, et al. Real-time epileptic seizure prediction using AR models and support vector machines [J]. IEEE Transactions on Biomedical Engineering, 2010, 57(5): 1124 - 1132.

[3] Laner M, Svoboda P, Rupp M. Parsimonious fitting of long-range dependent network traffic using ARMA models [J]. IEEE Communications Letters, 2013, 17(12): 2368 - 2371.

[4] Narendra B C, Eswara R B. A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data [J]. Applied Soft Computing, 2014, 23(1): 27 - 38.

[5] 孟庆芳, 陈珊珊, 陈月辉, 等. 基于递归量化分析与支持向量机的癫痫病脑电自动检测方法[J]. 物理学报, 2014, 63(5): 1 - 7.
(Meng Qing-fang, Chen Shan-shan, Chen Yue-hui, et al. Automatic detection of epileptic EEG based on recurrence quantification analysis and SVM [J]. Acta Physica Sinica, 2014, 63(5): 1 - 7.)

[6] 李琨, 韩莹, 黄海礁. 基于 IBH - LSSVM 的混沌时间序列预测及其在抽油井动液面短期预测中的应用[J]. 信息与控制, 2016, 45(2): 241 - 247.
(Li Kun, Han Ying, Huang Hai-jiao. Chaotic time series prediction based on IBH-LSSVM and its application to short-term prediction of dynamic fluid level of the oil wells [J]. Information and Control, 2016, 45(2): 241 - 247.)

[7] Du W, Leung S Y S, Kwong C K. Time series forecasting by neural networks: a knee point-based multi-objective evolutionary algorithm approach [J]. Expert Systems with Applications, 2014, 41(7): 8049 - 8061.

[8] Wang Z X, Pei L L. An optimized grey dynamic model for forecasting the output of high-tech industry in China [J]. Mathematical Problems in Engineering, 2014, 2014(1): 1 - 7.

[9] Jaeger H, Haas H. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication [J]. Science, 2004, 304(5667): 78 - 80.

[10] 田中大, 李树江, 王艳红, 等. 基于 KPCA 优化 ESN 的网络流量预测方法[J]. 电机与控制学报, 2015, 19(12): 114 - 120.
(Tian Zhong-da, Li Shu-jiang, Wang Yan-hong, et al. Network traffic prediction method based on KPCA optimized ESN [J]. Electric Machines and Control, 2015, 19(12): 114 - 120.)

[11] Hatamlou A. Black hole: a new heuristic optimization approach for data clustering [J]. Information Science, 2013, 222(1): 75 - 184.

[12] Li K, Gao X W, Zhou H B, et al. Fault diagnosis for down-hole conditions of sucker rod pumping systems based on the FBH-SC method [J]. Petroleum Science, 2015, 12(1): 135 - 147.

[13] 王通, 高宪文, 蒋子健. 基于黑洞算法的 LSSVM 的参数优化[J]. 东北大学学报(自然科学版), 2014, 35(2): 170 - 174.
(Wang Tong, Gao Xian-wen, Jiang Zi-jian. Parameters optimizing of LSSVM based on black hole algorithm [J]. Journal of Northeastern University(Natural Science), 2014, 35(2): 170 - 174.)

[14] Kim H S, Eykholt R, Salas J D. Nonlinear dynamics, delay times, and embedding windows [J]. Physica D: Nonlinear Phenomena, 1999, 127(1): 48 - 60.