物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于混沌理论和改进径向基函数神经网络的网络舆情预测方法

魏德志　陈福集　郑小雪

Internet public opinion chaotic prediction based on chaos theory and the improved radial basis function in neural networks

Wei De-Zhi　Chen Fu-Ji　Zheng Xiao-Xue

引用信息 Citation: Acta Physica Sinica, 64, 110503 (2015)　DOI: 10.7498/aps.64.110503
在线阅读 View online: http://dx.doi.org/10.7498/aps.64.110503
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2015/V64/I11

基于混沌理论和改进径向基函数神经网络的网络舆情预测方法*

魏德志[1)2)†] 陈福集[1)] 郑小雪[1)]

1) (福州大学经济与管理学院, 福州 350108)

2) (集美大学诚毅学院, 厦门 361021)

(2014 年 9 月 12 日收到; 2014 年 10 月 27 日收到修改稿)

网络舆情发展趋势具有混沌系统的特征, 提出一种基于EMPSO-RBF 神经网络的方法对网络舆情的发展趋势进行预测. 首先根据Lyapunov 指数证明网络舆情具备混沌的特征, 然后对网络舆情时间序列数据进行相空间重构, 最后采用EMPSO-RBF 方法进行预测, 并和其他模型进行对比试验, 实验结果表明EMPSO-RBF 方法具有较高精确度.

关键词: 神经网络, 混沌系统, 时间序列, 网络舆情

PACS: 05.45.Tp, 05.45.Gg **DOI:** 10.7498/aps.64.110503

1 引 言

在互联网的新时代, 无论是重大事件或国际活动, 在网络上短时间就可以形成舆论, 这些舆论可能会产生巨大力量. 面对几亿网民和几百万的媒体网站, 每天都能产生海量的网络舆情信息. 网络舆情趋势的发展由于广大网民的参与, 具有无规则和随机变化等特点, 对网络舆情发展趋势的准确预测有利于政府对舆情信息的监控和引导, 也有利于社会的和谐稳定.

目前, 可以用于网络舆情发展趋势预测分析的方法有最小均方算法差分[1]、混沌系统[2]、自回归移动平均(ARIMA)[3]、动力系统[4]、支持向量机[5−7]、神经网络等[8−14]. 由于网络舆情发展趋势的预测具有复杂性和非线性、采用传统统计学的方法具有一定的局限性, 机器学习方法中支持向量机和神经网络是目前用于非线性系统预测最主要的两种方法. 支持向量机是一种有关统计学的模型算法, 对非线性函数的逼近可以达到很好的效果, 但是对比神经网络类的算法在预测准确性没有特别的优势, 而且算法的参数选择太过于依赖人工. 人工神经网络经过长时间的训练, 对非线性函数可以达到任意精度. RBF 网络(径向基函数网络)和其他网络相比, 具有拓扑结构简单、学习快速高效, 是前向网络的一种主要结构, 广泛应用于实时自适应系统, 该网络经过证明只要节点数足够多, 可以以任何精度逼近单值函数, 因此得到了广泛应用. RBF 神经网络应用于预测的难点主要是本身网络结构的确定和相关参数的优化, 目前还没有一种方法可以完美解决这个问题.

由于网络舆情具有无规则、随机变化、非线性等特点, 具有一定的混沌特性, 以上方法应用于舆情趋势预测时并没有考到混沌特性, 为了提高预测精度, 本文首先结合混沌理论, 证明网络舆情发展趋势具有混沌性, 并采用相空间重构理论对预测模型进行建模, 然后对RBF 神经网络进行改进, 提出一种基于EM 聚类算法和改进的PSO 算法优化RBF 神经网络的混合算法(EMPSO-RBF)用于网络舆情发展趋势的预测.

* 国家自然科学基金 (批准号: 71271056) 和福建省教育厅项目 (批准号: C13001, JA14368) 资助的课题.

† 通信作者. E-mail: weidezhi@163.com

2 网络舆情发展趋势混沌特性分析及建模

混沌系统具有乱序、无规则、随机、对初始条件依赖等特点, 它是研究演化过程的一种科学系统. 网络舆情信息主要是由广大网民和网媒参与形成的媒体信息, 每个网民都有自己的喜好、习惯和观点, 参与每个舆情信息的用户数和网媒数无法预测, 用户具有无规则和随机性, 因此导致网络舆情信息具有混沌的特性, 采用混沌理论思想来对舆情信息分析有利于对未来发展趋势的预测.

2.1 舆情数据样本来源

百度公司开发的百度指数应用能统计某个关键词在一定时间段内被搜索的次数, 它以海量网民行为数据为基础的数据分享平台, 统计结果比较客观和全面. 选择“两会”作为舆情信息的关键字, 统计出从2014-5-1到2014-9-6每天的趋势指数, 具体如图1 所示. 把百度指数的时间序列数据作为网络舆情发展趋势的历史数据进行分析是比较可靠和准确的, 百度指数作为舆情数据样本的来源也是比较合理的.

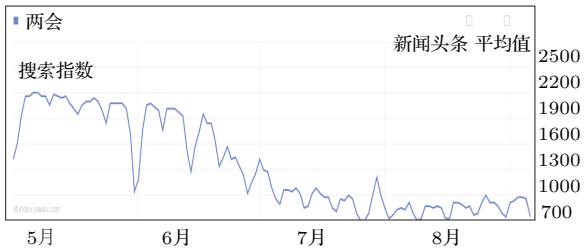

图1 “两会”百度指数趋势图

Fig. 1. Baidu index chart of “NPC and CPPCC”.

2.2 网络舆情发展趋势混沌性判断分析

系统行为是否有混沌特性可以从定性和定量分析两种方法进行判断. 定性方法主要是根据时间序列中的数据进行大体粗略分析, 从中找出混沌的特性, 定量方法有Lyapunov指数、关联维数等方法. Lyapunov指数方法用于混沌性的判断具有简单易于计算的特点, 本文采用该方法对网络舆情发展趋势的混沌特性进行证明. Lyapunov指数λ_1用于描述混沌系统初始敏感性, 可以计算出初始不确定性的指数发散放大率, 如果λ_1大于0, 则系统就具有混沌特性, 采用小数据量法求指数λ_1的计算步骤如下.

1) 对网络舆情发展趋势时间系列数据进行傅里叶变换并求出平均轨道周期P.

2) 利用C-C方法计算出延迟时间τ和嵌入维数m, 并重构相空间$\{Y_j, j=1,2,\cdots,M\}$.

3) 对每个点Y_j的最近临近点进行短暂分离限制, 计算邻点对应i个离散时间步长后的距离$d_j(i)$的公式为

$$d_j(i)=\left|Y_{j+i}-\hat{Y}_{j+i}\right|, \quad (i=1,2,\cdots,\min(M-j,M-\hat{j})). \tag{1}$$

4) 对每个i计算出所有j的$\ln d_j(i)$平均值$y(i)$的公式为

$$y(i)=\frac{1}{q\Delta t}\sum_{j=1}^{q}\ln d_j(i), \tag{2}$$

式中q是非零$d_j(i)$数目, 采用最小二乘法得到回归直线, Lyapunov指数λ_1就是该直线的斜率.

通过以上方法可以计算出“两会”网络舆情发展趋势时间序列的嵌入维数$m=8$和延迟时间$\tau=3$, 离散傅里叶平均周期$p=1$, 根据小数据量法计算出$\lambda_1=0.019$大于0, 结果表明网络舆情发展趋势时间序列具有混沌特性.

2.3 基于相空间重构的网络舆情发展趋势预测模型

对给定网络舆情发展趋势时间序列的数据集$D=\{x(t),t=1,2,\cdots,N\}$, N是给定时间序列的长度, 通过C-C方法计算出嵌入维数m和延迟时间τ, 采用延迟坐标法对相空间进行重构可以得到矩阵$\boldsymbol{X}$和$\boldsymbol{Y}$为

$$\boldsymbol{X}=\begin{pmatrix} x(1) & x(1+\tau) & \cdots & x(1+(m-1)\tau) \\ x(2) & x(2+\tau) & \cdots & x(2+(m-1)\tau) \\ \vdots & \vdots & \vdots & \vdots \\ x(M) & x(M+\tau) & \cdots & x(M+(m-1)\tau) \end{pmatrix}, \tag{3}$$

$$\boldsymbol{Y}=\begin{pmatrix} x(2+(m-1)\tau) \\ x(3+(m-1)\tau) \\ \vdots \\ x(M+1+(m-1)\tau) \end{pmatrix}. \tag{4}$$

根据Takens理论, 重构相空间轨迹与原系统在同胚意义下动力学等价, 舆情信息预测模型就是

根据相空间中的点 $X(t)$ 预测出 $Y(t)$, 从中找到一个映射函数 f, 具体如下式, 预测出下一个时间序列的数据点:

$$Y(t) = f(x(t)). \tag{5}$$

该预测模型具有混沌特性, 映射函数 f 是一个非线性结构, 传统统计学方法如自回归、滑动平均、ARIMA 并不适合求解, 因此本文采用机器学习方法 RBF 神经网络的方法进行求解, 为了提高预测精确度, 对 RBF 神经网络的方法进行改进, 更好的适合网络舆情发展趋势的预测.

3 RBF 神经网络

RBF 网络是一种三层结构的神经网络: 第一层是输入层, 该层节点个数主要由输入向量 x 的维数决定; 第二层是隐含层, 和输入节点相连接, 该层节点主要由训练数据点的个数决定; 第三层是输出层, 输出数据的维数等于该层的节点数. 假设有 n 个训练样本, 则 RBF 网络结构如图 2 所示.

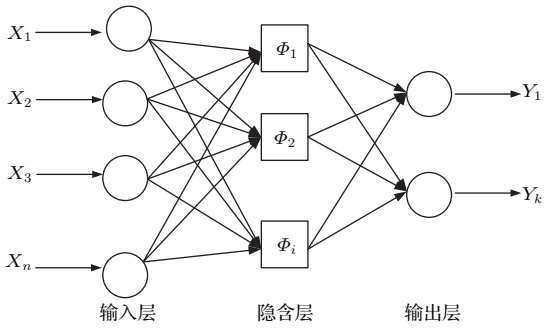

图 2 RBF 神经网络结构图

Fig. 2. Structure of RBF neural network.

设输入 $X_n = [x_{n1}, x_{n2}, \cdots, x_{nm}]$, 实际输出为 $Y_k = [y_{k1}, y_{k2}, \cdots, y_{kj}]$, 那么从输入到输出的函数映射公式为

$$y_{kj} = \theta_j + \sum_{i=1}^{I} \omega_{ij} \phi(X_n, X_i),$$
$$j = 1, 2, \cdots, J, \tag{6}$$

$$\phi(X_n, X_i) = \exp\left(-\frac{1}{2\sigma^2} \|X_n - X_i\|^2\right). \tag{7}$$

(6) 式中, n 为输入样本的个数, J 为输出单元的个数, I 为隐含层的节点数, $X_i = [x_{i1}, x_{i2}, \cdots, x_{im}]$ 为基函数的中心, θ_j 为网络阈值, ω_{ij} 为网络连接权值. (7) 式中, 基函数 ϕ 采用高斯函数, σ 为基宽向量. RBF 网络的精度主要由中心节点、基宽向量、阈值、网络值等参数决定, 通过对网络参数进行优化, 可以提高网络的逼近性能.

4 改进的神经网络算法 EMPSO-RBF

4.1 EM 聚类算法结合高斯模型改进径向基函数

传统聚类算法比如 K-Means 等对 RBF 的优化通过对网络结构的优化来实现, 具体思想是从网络节点个数、中心点个数来提高网络的精确度, 这些方法从网络的外在特性进行分析改进, 但是对内在的神经元的宽度以及径向基函数的改进就基本上没有. (7) 式中基宽向量 σ 可以采用协方差的形式表示为 $\Sigma_i = \mathrm{diag}\{\sigma_i^2, \sigma_i^2, \ldots, \sigma_i^2\}$ 或者 $\sigma_i^2 \times \mathrm{diag}\{1, 1, \cdots, 1\}$. 传统高斯混合模型中 $f_i(x)$ 的公式为

$$f_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \times \exp\left[-\frac{1}{2}(x - u_i)^{\mathrm{T}} (\Sigma_i)^{-1} \times (x - u_i)\right]. \tag{8}$$

比较 (7) 和 (8) 式可以发现, 两者的差别主要体现在协方差 Σ_i 上, 其中 (8) 式中 $\Sigma_i = \mathrm{diag}\{a_{11}, a_{22}, \cdots, a_{d,d}\}$, 其中 $a_{11} \neq a_{22} \neq \cdots \neq a_{d,d}$, 系数不完全性等, 但是 (7) 式中的协方差系数是完全相等的即 $\sigma_i^2 \times \mathrm{diag}\{1, 1, \cdots, 1\}$. (7) 式中协方差形式表达的信息量小, 形式简单, 无法根据样本的数据进行类别划分, 体现数据的分布情况. 但是如果使用 $\Sigma_i = \mathrm{diag}\{a_{11}, a_{22}, \cdots, a_{d,d}\}$, 其中 $a_{11} \neq a_{22} \neq \cdots \neq a_{d,d}$, 每个参数都不完全相等, 可以体现出每个聚类对象之间的不同, 信息量大, 更好统计出数据的分布情况. 因此将隐含层径向基函数修改为

$$\psi(X_n, u_i) = \exp\left(-\frac{1}{2}(X_n - u_i)^{\mathrm{T}} (\Sigma_i)^{-1} \times (X_n - u_i)\right),$$
$$i = 1, \cdots, M. \tag{9}$$

最大似然估计算法简称 EM 算法, 是一种以统计学为基础、能抗噪音干扰、具有鲁棒性的聚类算法, 可以快速有效的计算出 (9) 式中的中心值 u_i 和方差 Σ_i, 并且把结果直接应用到改进隐含层径向基函数 $\psi(X_n, u_i)$ 中, 减少了聚类计算的次数, 根据

样本的概率值来进行数据对象的分割计算, 抗噪音干扰性强, 可以进行高维空间计算, 算法初始化容易并且收敛快速.

4.2 结合PSO算法进行优化

PSO算法是一种模拟鸟类捕食行为的群体智能进化算法, 每个粒子代表问题的一个解, 粒子在空间运动中根据适应度函数计算结果不断进行位置和速度的调整, 最终在空间中寻找到最优解. 该算法具有容易收敛, 简单易操作等特点, 特别适合于对RBF神经网络的优化, 如果把RBF神经网络看成是一个预测函数, PSO算法优化RBF神经网络相当于优化预测函数中的参数, 优化后的RBF神经网络的预测效益优于未优化的RBF神经网络[13]. 但是PSO算法存在着收敛早熟、有一定误差以及不容易进行全局寻最优等缺点, 需要进一步改进以适应RBF神经网络的优化.

4.3 混合算法的编码

首先利用EM聚类算法求出对应的中心值u_i和方差Σ_i, 得到RBF径向基函数, 确定网络基本结构. 然后将需要优化的网络的隐含层中心值、方差、阈值、权值参数的初值统一整合到向量X中, 作为混合算法寻优的位置向量, 编码使用实数. 混合算法当中的粒子编码顺序具体为

$$[x_{n1}, x_{n2}, \cdots, x_{nm}] \,||\, [v_{n1}, v_{n2}, \cdots, v_{nm}] \,||\, f\,[x_{n1}, x_{n2}, \cdots, x_{nm}],$$

分别代表粒子位置、粒子速度、目标函数.

4.4 混合算法的非线性惯性权重

惯性权重ω代表的是当前粒子以多大的程度继承上一代粒子的速度, Shi提出了线性递减惯性权重来更好平衡算法的全局和局部搜索能力, 具体如公式为

$$\omega(k) = \omega_{\text{start}} - (\omega_{\text{start}} - \omega_{\text{end}}) \times \frac{(k_{\max} - k)}{k_{\max}}. \tag{10}$$

为了解决线性递减惯性权重不能很好处理非线性问题, 本文对(10)式进行改进, 提出一种非线性惯性权重, 具体公式为

$$\omega(k) = \omega_{\text{end}} + (\omega_{\text{start}} - \omega_{\text{end}}) \times \exp\left(-12\frac{k}{k_{\max}}\right)^2. \tag{11}$$

其中ω_{start}代表开始惯性权重、ω_{end}代表结束惯性权重、$k_{\max}$代表最大迭代次数、k代表当前迭代次数、$\omega(k)$代表当前惯性权重.

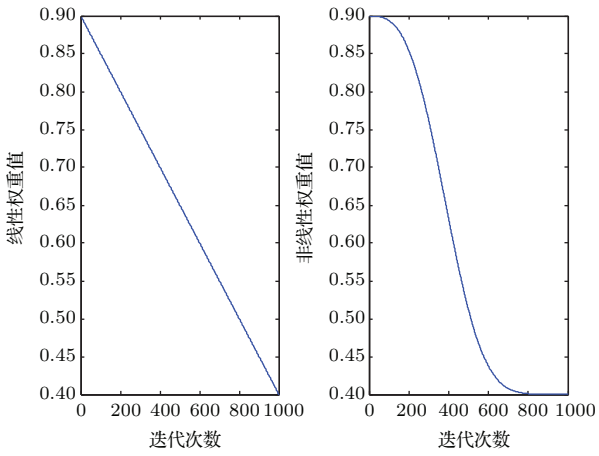

图3 惯性权重ω变化过程

Fig. 3. Change process of the inertia weight.

取值$\omega_{\text{start}} = 0.9$, $\omega_{\text{end}} = 0.4$, 从图3中可以看出惯性权重$\omega$根据(10)和(11)式迭代变化的整个过程, (11)式变化过程比线性结构更加平滑, 在变化初期惯性权重取值较大对全局搜索比较有利, 在变化后期惯性权重取值较小对局部搜索有利, 加快算法的收敛.

4.5 混合算法的变异

目前部分学者把遗传算法中变异的操作引入PSO算法进行改进, 主要是根据一定的概率在粒子更新的时候进行变异操作, 并没有考虑到粒子群在整个迭代过程中粒子状态的变化. 本文提出一种分段变异算子, 在算法迭代初期对适应度较低的粒子进行变异, 把该粒子的位置更新为适应度高粒子的平均值; 在后期预设一个阈值, 对个体和全局极值进行随机扰动, 具体扰动公式为

$$V_i^{k+1} = \omega V_i^k + c_1 r_1 \left(r_3 P_i^k - X_i^k\right) + c_2 r_2 \left(r_4 P_{\text{g}}^k - X_i^k\right), \tag{12}$$

其中V_i代表粒子速度, P_i代表个体极值, P_{g}代表全局极值, V_i代表粒子位置, c_1, c_2, r_1, r_2, r_3, r_4为具体给定常数.

随机扰动的基本思想为: 如果大于阈值, 按照(12)式进行更新, 否则按照原公式进行更新计算. 通过以上的变异操作, 增强粒子群的多样性, 拓展

了种群的空间搜索能力, 防止出现收敛过早的现象, 更好得到全局最优解.

4.6 混合算法具体步骤

1) 给定神经网络的训练数据集, 包含输入和输出样本集.

2) 利用EM聚类算法确定RBF神经网络的中心值 u_i 和方差 Σ_i, 并确定径向基函数.

3) 对网络的隐含层中心值、方差、阈值、权值参数进行实数编码, 随机产生粒子种群, 同时初始化PSO参数.

4) 迭代更新粒子的速度和位置.

5) 确定PSO算法的适应度函数, 采用输出值的均方根误差公式作为适应度的计算公式, 具体公式为

$$e=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}, \tag{13}$$

其中 e 代表均方根误差、$\hat{y}$ 为预测值、y_i 为真实值、n 为样本数. 根据误差适应度函数计算结果对个体和群体极值进行更新.

6) 判断是否达到算法结束条件, 结束条件为: 达到设定的群体极值或迭代到设定的代数, 具体以那个条件先到为先. 如果不达到转到第4), 达到得到一组最优的网络参数.

7) 利用RBF网络算法进行实验数据训练预测, 得到预测结果.

混合算法充分利用了EM聚类算法和改进PSO的优点, 对RBF神经网络进行优化和改进, 分别从前期初始化网络结构和后期优化网络参数两个角度进行. 首先采用EM聚类算法获得中心值 u_i 和方差 Σ_i, 结合传统高斯模型对径向基函数进行改进; 然后采用改进PSO进化算法根据误差不断对网络参数进行优化得到相关网络参数, 提高逼近精度, 更好解决网络舆情发展趋势非线性系统预测的问题.

5 仿真实验

5.1 实验条件

本文的实验仿真环境为Matlab2010b, 机器为Dell, CPU: i5 3.2G, RAM: 4G, 操作系统为Win-xp. 通过神经网络工具箱编写EMPSO-RBF神经网络预测算法, 对“ 两会”等网络舆情发展趋势时间序列数据进行预测实验, 最后和参考文献中的其他预测算法进行误差指标对比. 实验中, 作为预测实验误差评价采用绝对误差Err为

$$\mathrm{Err}=\hat{y}_i-y_i; \tag{14}$$

均方根误差Rmse为

$$\mathrm{Rmse}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}; \tag{15}$$

相对误差Perr为

$$\mathrm{Perr}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}y_i^2}; \tag{16}$$

平均绝对百分率误差Smape为

$$\mathrm{Smape}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i+\hat{y}_i}\right|. \tag{17}$$

式中 $\hat{y}$ 为预测值, y_i 为真实值, n 为样本数, 误差Rmse代表预测值和真实值之间的平均偏离程度, 误差Perr代表绝对误差和真实值之间的比值, 误差Smape代表评估时间序列数据的趋势, 衡量拟合的准确程度.

5.2 实验结果

将“两会”等网络舆情发展趋势时间序列数据从2014年5月1日到9月1日产生的124个分量的时间序列数据分成两种数据样本, 分别是训练和预测. 其中前94个数据作为训练样本, 后30个数据作为预测样本. 图4给出了不同算法在“两会”网络舆情发展趋势在30个预测样本中的预测实验结果分析, 图5给出了绝对误差结果的比较分析, 表1给出了混合算法EMPSO-RBF和文献其他算法在误差Rmse, Perr和Smape的比较结果.

从图3可以看出本文算法在对真实值预测中表现比较好, 能比较真实反映实际值的变化趋势; 从图4中可以看出本文算法的绝对误差比其他算法明显低, 预测效果最好. 从表2 中可以看出不同模型算法在误差指标RMSE, Perr, Smape中的表现情况, 在误差Perr中本文的算法比其他模型算法的误差指标提高了一个数量级左右, 在RMSE和Smape误差表现中也是所有算法中最好的, 在误差精度控制得到较好的效果. 在预测时间比较上, 本文的速度也是比较快, 运行时间相对其他算法较短, 效率较高.

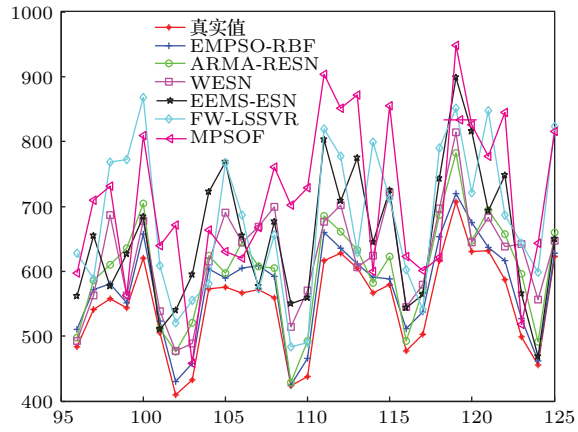

图 4 (网刊彩色)“两会”舆情信息实验结果比较图

Fig. 4. (color online) Experimental results comparison chart of the public opinion information of “NPC and CPPCC”.

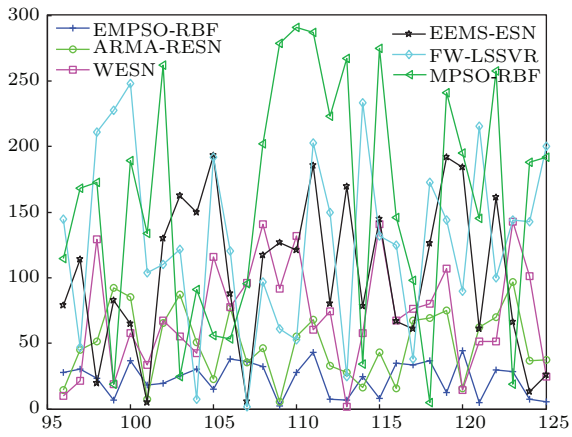

图 5 (网刊彩色) 绝对误差比较分析图

Fig. 5. (color online) Comparison chart of absolute error.

表 1 不同模型算法对“两会”舆情信息预测误差表

Table 1. Error prediction of different model about the public opinion information of “NPC and CPPCC”.

模型算法	RMSE	Perr	Smape	预测时间/s
本文 EMPSO-RBF	26.4085	6.9368×10^{-5}	2.0757×10^{-2}	0.823
ARMA-RESN [9]	55.6963	2.8211×10^{-4}	4.2626×10^{-2}	1.607
WESN [10]	82.2600	5.7337×10^{-4}	6.1237×10^{-2}	1.418
EEMD-ESN [11]	117.1652	1.0468×10^{-3}	8.3442×10^{-2}	1.801
FW-LSSVR [12]	145.4406	1.4858×10^{-3}	1.0175×10^{-1}	1.472
MPSO-RBF [13]	181.5077	2.1281×10^{-3}	1.2183×10^{-1}	0.979

5.3 其他混沌信息预测

为了进一步验证该模型的性能, 本文采用传统 Henon 非线性混沌系统对该模型进行验证. Henon 系统的参数初始值为 $a = 1.4$, $b = 0.3$ [14], 系统产生的混沌系列取 x 分量作为实验数据. x 序列产生的前 12000 点舍弃, 取后面 3000 点作为训练实验数据. 将混沌系统产生的后 3000 个 x 分量的时间序列数据分成两种数据样本, 分别是训练和预测. 其中前 2000 个数据作为训练样本, 后 1000 个数据作为预测样本. 表 2 给出了混合算法和文献其他算法在误差 Rmse, Perr 和 Smape 以及预测时间的比较结果.

最后通过百度指数, 分别收集关键词为“反腐”, “中国梦”, “世界杯”, “食品安全”四个热点舆情话题的历史时间序列数据进一步实际情况验证, 收集时间为 2014 年 5 月 1 日到 9 月 1 日. 表 3 给出了本文混合算法的误差表现和预测时间的情况.

表 2 不同模型算法对 Henon 系统预测误差表

Table 2. Error prediction of different model about the Henon.

模型算法	RMSE	Perr	Smape	预测时间/s
本文 EMPSO-RBF	2.1134×10^{-3}	8.5319×10^{-6}	5.6213×10^{-3}	2.365
ARMA-RESN [9]	1.2234×10^{-2}	4.3156×10^{-5}	2.3654×10^{-2}	4.291
WESN [10]	2.6734×10^{-2}	9.6372×10^{-4}	8.1564×10^{-2}	4.372
EEMD-ESN [11]	5.7047×10^{-2}	3.4653×10^{-3}	7.6985×10^{-2}	4.378
FW-LSSVR [12]	7.670×10^{-2}	5.8946×10^{-3}	1.3659×10^{-1}	3.422
MPSO-RBF [13]	9.6423×10^{-2}	3.1428×10^{-4}	1.8623×10^{-1}	2.413

表 3 EMPSO-RBF 针对不同话题的预测误差

Table 3. The prediction errors of different topics by EMPSO-RBF.

舆情话题	RMSE	Perr	Smape	预测时间/s
反腐	27.3871	2.4994×10^{-6}	3.9243×10^{-3}	0.823
中国梦	36.2356	5.6874×10^{-5}	1.568×10^{-2}	0.856
世界杯	29.6812	8.6329×10^{-6}	8.6954×10^{-3}	0.796
食品安全	25.5234	5.6987×10^{-6}	6.2531×10^{-3}	0.756

从表 2 和表 3 可以看出 EMPSO-RBF 在针对不同混沌时间序列中获得较好的预测精确度, 预测的各项误差偏差不大, 没有出现特别的异常现象; 预测时间也是比较正常, 没有出现大的波动, 说明算法的收敛性是比较可靠的. 因此基于混沌理论的 EMPSO-RBF 算法是一种预测效果好, 通用性比较强的网络舆情预测模型.

6 结 论

网络舆情信息受到网络众多网民和网媒的影响, 信息具有无规则、随机等混沌特性, 是一种非线性的复杂演化系统, 传统基于统计和机器学习的方法很难建立对应的模型并进行有效的预测. 针对网络舆情的混沌特性, 本文首先结合混沌理论思想, 通过相空间重构理论对网络舆情发展趋势预测进行建模, 然后提出了一种基于 EM 算法和改进的 PSO 算法优化 RBF 神经网络的混合算法 EMPSO-RBF 来对模型进行求解, 通过对不同混沌信息时间序列实验验证, 本文模型算法能够比较精确的模拟混沌信息的时间序列, 更好描述不同网络舆情信息的发展趋势, 预测结果利于政府对舆情信息的监控和引导, 也有利于社会的和谐稳定.

参考文献

[1] Bilal S, Ijaz M Q, Ihsanulhaq, Shafqatullah 2014 *Chin. Phys. B* **23** 030502

[2] Zhang Y 2013 *Chin. Phys. B* **22** 050502

[3] Zhou Y M, Li B C 2013 *Journal of Data Acquisition & Processing.* **28** 69 (in Chinese) [周耀明, 李弼程 2013 数据采集与处理 **28** 69]

[4] Liu J, Shi S T, Zhao J C 2013 *Chin. Phys. B* **22** 010505

[5] Sun Z H, Jiang F 2010 *Chin. Phys. B* **19** 110502

[6] Keerthi S S, Lin C J 2003 *Neural Computation.* **15** 1667

[7] Amjady N, Keynia F 2008 *Int. J. Elec. Power.* **30** 533

[8] Wu X D, Wang Y L, Liu W T, Zhu Z Y 2011 *Chin. Phys. B* **20** 069201

[9] Tang Z J, Ren F, Peng T, Wang W B 2014 *Acta Phys. Sin.* **63** 050505 (in Chinese) [唐舟进, 任峰, 彭涛, 王文博 2014 物理学报 **63** 050505]

[10] Song T, Li J 2012 *Acta Phys. Sin.* **61** 080506 (in Chinese) [宋彤, 李菡 2012 物理学报 **61** 080506]

[11] Zhang X Q, Liang J 2013 *Acta Phys. Sin.* **62** 050505 (in Chinese) [张学清, 梁军 2013 物理学报 **62** 050505]

[12] Zhao Y P, Zhang L Y, Li D C, Wang L F, Jiang H Z 2013 *Acta Phys. Sin.* **62** 120511 (in Chinese) [赵永平, 张丽艳, 李德才, 王立峰, 蒋洪章 2013 物理学报 **62** 120511]

[13] Xiao B X, Wang X W, Liu Y F 2007 *Journal of System Simulation.* **19** 1382 (in Chinese) [肖本贤, 王晓伟, 刘一福 2007 系统仿真学报 **19** 1382]

[14] Henon M 1976 *Commun. Math. Phys.* **5** 5069

Internet public opinion chaotic prediction based on chaos theory and the improved radial basis function in neural networks*

Wei De-Zhi[1)2)†] Chen Fu-Ji[1)] Zheng Xiao-Xue[1)]

1) (*School of Economics and Management, Fuzhou University, Fuzhou 350108, China*)

2) (*Jimei University Chengyi College, Xiamen 361021, China*)

(Received 12 September 2014; revised manuscript received 27 October 2014)

Abstract

Information of internet public opinion is influenced by many netizens and net medias; characteristics of this information are non regular, stochastic, and may be expressed by a nonlinear complex evolution system. Corresponding model is difficult to establish and effectively predicted using the traditional methods based on statistical and machine learning. Characteristics of internet public opinion are chaotic, so the chaos theory can be introduced to research first, then the information of internet public opinion having chaotic characteristic is proved by the Lyapunov index. The model to predict the development trend of internet public opinion is next established by the phase space reconstruction theory. Finally, the hybrid algorithm EMPSO-RBF which is based on EM algorithm and the RBF neural network optimized by the improved PSO algorithm is proposed to solve the model. The hybrid algorithm fully takes the advantage of the EM clustering algorithm and the improved PSO, so the RBF neural network is improved by initializing the network structure in the early stage and optimizing the network parameters later. First, the EM clustering algorithm is used to obtain the center value and variance, and the radial basis function is improved with the combination of traditional Gauss model. Then the relevant network parameters are obtained by the improved PSO algorithm which is based on error optimizing the network parameters constantly. The model algorithm can be accurately simulated in the time series of chaotic information by experiments which are validated by different chaotic time series information; and it can better describe the development trend of different information of internet public opinion. The predicted results are made for government to monitor and guide the information of internet public opinion and benefit the social harmony and stability.

Keywords: neural network, chaotic system, time series, internet public opinion

PACS: 05.45.Tp, 05.45.Gg **DOI:** 10.7498/aps.64.110503

* Project supported by the National Natural Science Foundation of China (Grant No. 71271056), and the Department of Education of Fujian Province, China(Grant Nos. C13001, JA14368).

† Corresponding author. E-mail: weidezhi@163.com