

基于变结构神经网络控制的决策树预测算法

单冬红,吕琼帅,赵伟艇

(平顶山学院 计算机学院(软件学院),河南 平顶山 467000)

摘 要: 为了提高统计数据时间序列准确预测的能力,针对当前的线性指数预测方法计算开销过大,且容易陷入局部最优的问题,提出一种基于变结构神经网络控制的决策树预测算法,对待处理的统计数据时间序列进行非线性特征相空间重构,在重构的高维特征相空间中进行决策树训练,提取统计数据信息流的主成分特征,对提取的特征参量采用变结构神经网络控制器进行伴随跟踪训练,进行数据聚类中心搜索和预测指向性误差的小扰动抑制,实现统计数据时间序列的准确预测。仿真结果表明,采用该算法进行统计数据预测的精度较高,收敛性较好,对数据信息流的走势拟合性能较好,性能优于传统方法。

关键词: 变结构神经网络;统计数据;决策树;预测;聚类

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2017)02—0092—05

Decision Tree Prediction Algorithm Based on Variable Structure Neural Network Control

SHAN Dong-hong, LV Qiong-shuai, ZHAO Wei-ting

(Department of Computer, Pingdingshan University, Pingdingshan 467000, China)

Abstract: In order to improve the ability to predict the time sequence statistics, excessive computational overhead for the linear current index prediction method, and easy to fall into local optimum, this paper proposes a prediction algorithm of decision tree structure based on neural network control, treat the statistical data of time series nonlinear characteristics of phase space reconstruction, decision making tree training in high dimensional feature reconstruction of phase space, principal component feature extraction of statistical data on the flow of information, the extracted characteristic parameters of the variable structure neural network controller with tracking training, data clustering search and forecast the pointing error of small disturbance suppression, realize the accurate prediction of statistical data of time series. The simulation results show that the algorithm has higher precision and better convergence performance, and the performance of the data flow is better than that of the traditional method.

Key words: variable structure neural network; statistical data; decision tree; prediction; clustering

1 引言

随着网络信息技术的不断深入发展,人类进入大数据信息处理时代,采用并行处理算法能够对具有海量数据规模、快速流转的数据流进行分布式处理、分布式数据库存储和管理,并提高数据分析的能力。对经济指数数据、交通流数据、故障监测数据、环境检测数据等统计数据信息流的处理和分析作为信

息处理领域的一个重要分支,对该类数据进行有效预测,在经济调控、交通调度和环境监测等领域具有广泛的应用价值^[1]。

统计数据预测采用分布式处理(Distributed Computing)、并行处理(Parallel Computing)和网格计算(Grid Computing)等方法,结合相关的数据特征挖掘和聚类,实现数据预测,对统计数据时间序列准确预测的影响因素众多,由于在网络空间和现实

生活中采集的统计数据时间序列具有时变性和随机性变化特点,对统计数据预测的准确性受到影响.传统方法中,对统计数据预测方法主要有 Wolf 一步预测法、AR 模型法、时频分析方法等^[2-3],上述方法属于线性时间序列分析方法,通过特征提取实现数据预测,进行统计数据指数时间序列的统计特征分析,取得了较好研究成果,其中,Becka 和 Vajtersic 提出的并行双端块雅克比(parallel two-side block-Jacobi)预测算法^[4],采用高阶谱特征提取可进行统计数据信息流的关键特征分解,构建端块雅克比矩阵实现统计数据的多层次分布式挖掘,提高在数据预测过程中的并行处理效率,但是该方法随着数据规模的扩大,计算开销增长较快,影响对数据走势预测的时效性;文献[5]提出一种基于高斯随机游走和个体筛选策略的统计数据差分进化预测算法,提取统计数据信息流的时延尺度特征参量,结合模糊聚类实现数据预测和特征挖掘,但是该算法的计算开销较大,且对初始值敏感,容易陷入局部收敛,数据预测的总体收敛性不好^[6].针对上述问题,本文提出一种基于变结构神经网络控制的决策树预测算法,首先对待处理的统计数据时间序列进行非线性特征相空间重构,在重构的高维特征相空间中进行决策树训练,实现对统计数据信息流的主成分特征提取,然后对提取的特征参量采用变结构神经网络控制器进行伴随跟踪训练,进行数据聚类中心搜索和预测指向性误差的小扰动抑制,实现统计数据时间序列的准确预测.最后进行仿真测试,得出有效性结论.

2 统计数据非线性特征相空间重构及预处理

2.1 统计数据非线性特征相空间重构

为了实现对统计数据信息流的优化准确预测,采用非线性时间序列分析方法构建统计数据的信息流模型,在统计数据信息样本库中,通过前期的统计分析和数据采样,得到统计数据信息流的单变量时间序列表示 $\{x_n\}$,其中数据采样样本长度为 N ,通过统计特征采样得到的统计数据时间序列的嵌入维数 m 、统计时延 τ ,构建一个 ARMA 模型表示统计数据时间序列的信息流模型为^[7]

$$x_n = x(t_0 + n\Delta t) = h[z(t_0 + n\Delta t)] + \omega_n \quad (1)$$

式中, $h(\cdot)$ 为统计数据时间序列的分布相轨迹, ω_n 为统计数据采样的统计误差. 对前期监测测量的统计数据时间序列 $\{x(t_0 + i\Delta t)\}$, $i = 0, 1, \dots, N-1$ 进行样本特征分布重组,把数据分布的特征轨迹映

射到高维特征空间中,相空间重构轨迹为

$$X = [s_1, s_2, \dots, s_K]_n = (x_n, x_{n-\tau}, \dots, x_{n-(m-1)\tau}) \quad (2)$$

式中, $K = N - (m-1)\tau$, 表示统计数据时间序列的数据相似度(相关性), m 为在相空间中的嵌入维数, $s_i = (x_i, x_{i+\tau}, \dots, x_{i+(m-1)\tau})^T$ 为数据预测过程中的伴随跟踪向量.

在相空间中,采用二维 C4.5 决策树模型进行统计特征预测和概率置信度分析,当相轨迹满足统计数据预测并行计算的特征泛函约束条件,计算 $\{\delta x_i(j_k) = x_{jk} - x_i \mid k \in 1, 2, \dots, N_b\}$, 产生 N_b 个 m 维向量,通过 QR 分解^[8],将重构相空间的时滞向量初始解 $X^{(0)}$ 划分为 N 部分,为 $x^{(1)}, x^{(2)}, \dots, x^{(N)}$, 即 $x^{(0)} = \bigcup_{i=1}^N x^{(i)}$, 通过自相关特征匹配,得到多层次分布的统计数据的最大 Lyapunov 指数谱筛选的迭代式为:

$$\max_{x_{a,b,d,p}} \sum_{a \in A} \sum_{b \in B} \sum_{d \in D} \sum_{p \in P} x_{a,b,d,p} V_p \quad (3)$$

$$\text{s. t. } \sum_{a \in A} \sum_{d \in D} \sum_{p \in P} x_{a,b,d,p} R_p^{bw} \leq K_b^{bw}(S), b \in B \quad (4)$$

在二维主特征空间中,进行数据集的残差数据合并和容易特征提取,采用决策数据模型进行伴随跟踪测试和数据分布向量的状态识别,假设统计数据时间序列采样的先验数据表示为 $\{x_i\}_{i=1}^N$,为了去除时间序列中的残差向量,通过数据特征伴随状态验证,得到具有相同的均值、方差的统计数据时间序列标量时间序列的特征相空间重构映射:

$$x_n = (x_n, x_{n-\tau}, \dots, x_{n-(m-1)\tau}) \quad (5)$$

数据流量序列的概率密度函数表达为

$$F_Y(y; \alpha, \lambda) = \int_0^y K_v(z) \frac{4}{\sqrt{\lambda} \Gamma(\alpha)} \left(\frac{t}{\sqrt{\lambda}}\right)^\alpha K_{\alpha-1}\left(\frac{2t}{\sqrt{\lambda}}\right) dt \quad (6)$$

式中, $K_v(z)$ 函数表示统计数据特征分布服从参数为 (α, β) 的均匀正态分布. 通过上述处理进行统计数据时间序列的非线性特征相空间重构,通过状态提取和模型提取,采用决策训练方法进行统计数据预测,统计数据信息流预测的总体实现流程如图 1 所示.

2.2 主成分特征提取预处理

在重构的高维特征相空间中进行决策树训练,实现对统计数据信息流的主成分特征提取,假设数据非线性特征相空间中 $x_n \rightarrow x_{n+1}$ 的演化数据集的决策树模型用四元组 (E_i, E_j, d, t) 来表达,其中 E_i ,

树中 i 级分叉树节点中对数据信息流预测的扰动误差向量为

$$\text{net}_i(k) = r_i(k) \quad i = 1, 2 \tag{17}$$

对于神经网络的第 j 个神经元, 输出量 $x_1, x_2, \dots, x_n$ 分别乘上权重值 $w_{1j}, w_{2j}, \dots, w_{nj}$, 对统计数据时间序列的主成分特征量进行预测误差扰动抑制, 得到预测估计值为

$$\hat{S}_w = \sum_{i=1}^c p_i \frac{1}{n_i} \sum_{k=1}^{n_i} [(\vec{X}_k^{(i)} - \vec{m}_i)(\vec{X}_k^{(i)} - \vec{m}_i)^T] \tag{18}$$

对数据预测的输出特征量按降序顺序排列为

$$S(\vec{X}_1) \geq S(\vec{X}_2) \geq \dots \geq s(\vec{X}_l) \tag{19}$$

通过上述采用变结构神经网络控制方法对数据聚类中心搜索和预测指向性误差的小扰动抑制, 提高了预测精度, 降低了预测误差。

4 仿真实验与结果分析

为了测试本文设计算法在实现对统计数据时间指数序列准确预测中的应用性能, 进行仿真实验, 分别取其中的 6 组数据样本进行数据预测分析, 神经网络的学习训练参数为 $NE_{j^*}(t) = 1 + 9e^{-t/1.024} = 12$, 隐含层神经元的初始学习速率 $\eta = 0.7$, 训练函数选择为 $f_1(x) = \sum_{i=1}^n x_i^2$, PID 神经网络隐含层节点数 $h = 6, 8, 10$, 数据统计信息特征采样周期为 10 Day, 迭代次数为 10 000 次, 根据上述仿真环境和参数设定, 进行数据的预测仿真分析, 得到原始采样的数据信息流时域波形如图 2 所示。

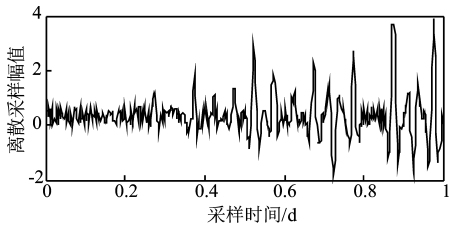


图 2 数据信息统计特征采样时域波形

以上述采样数据为测试集, 进行数据流发展指数预测, 采用本文算法进行决策树训练和主特征量提取, 分离残差数据, 得到分离结果如图 3 所示。

分析图 3 结果得知, 采用本文方法进行数据预测, 能有效实现对残差干扰数据分离, 提高了数据预测中的主成分特征提取精度, 保障了预测准确性, 为了对比算法性能, 采用本文算法和传统算法进行预测的误差对比结果如图 4 所示。不同算法的预测时间开销和精度提升对比见表 1。

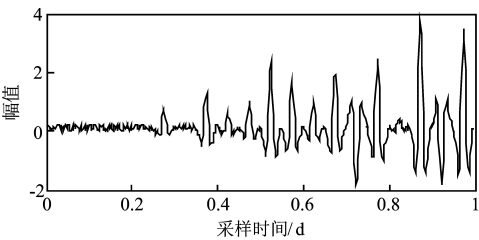


图 3 数据信息流的特征成分特征提取输出

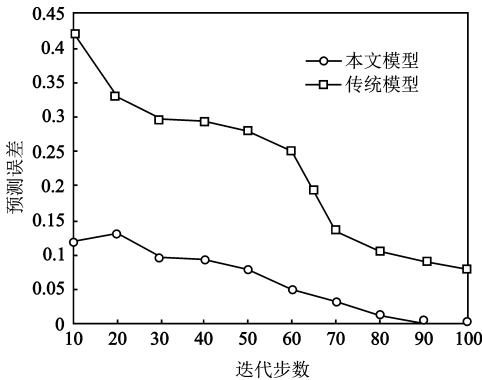


图 4 预测误差对比

表 1 不同算法的统计数据预测的性能对比

算法	内存开销/GB	均方误差 RMSE	运行时间/s	精度提升
Wolf 一步预测	420	0.143 2	0.354	—
AR 模型	510	0.048 6	3.453	—
时频分析	300	0.065 4	5.586	—
本文方法	120	0.000 343 2	0.655 4	22.301%

分析上述图 4 和表 1 的结果得知, 采用本文算法进行统计数据指数预测的内存开销最小, 本文方法的预测误差和运行时间都较小, 采用时频分析方法的计算开销较大, 采用 Wolf 一步预测算法的预测误差最大, 这是因为迭代次数太少的原因, 但是该方法的时间开销最少, 本文方法在精度和时间开销方面取得权衡, 通过 1 000 次实验取平均值, 得到本文方法的预测精度提升 22.31%, 采用本方法进行数据预测, 能较好地拟合数据信息流的发展趋势, 预测性能提高显著。

5 结束语

本文研究了统计数据指数时间序列的准确预测问题, 提出一种基于变结构神经网络控制的决策树预测算法, 对待处理的统计数据时间序列进行非线性特征相空间重构, 在重构的高维特征相空间中进行决策树训练, 对统计数据信息流的主成分特征提

取,对提取的特征参量采用变结构神经网络控制器进行伴随跟踪训练,实现数据聚类中心搜索和预测指向性误差的小扰动抑制,完成统计数据时间序列的准确预测.研究得知,采用该算法进行统计数据预测的精度较高,误差较小,内存开销和计算开销等指标参量都具有一定的优越性,展示了较高的应用价值.

参考文献:

[1] 陆兴华,陈平华. 基于定量递归联合熵特征重构的缓冲区流量预测算法[J]. 计算机科学, 2015, 42(4): 68-71.

[2] 杨雷,李贵鹏,张萍. 改进的 Wolf 一步预测的网络异常流量检测[J]. 科技通报, 2014, 30(2): 47-49.

[3] 梁聪刚,王鸿章. 微分进化算法的优化研究及其在聚类分析中的应用[J]. 现代电子技术, 2016, 39(13): 103-107.

[4] Ortiz A, Gorriz J M, Ramirez J, et al. Improving MR brain image segmentation using self-organising maps and entropy-gradient clustering[J]. Information Sciences, 2014, 262(3): 117-136.

[5] 李牧东,赵辉,翁兴伟,等. 基于最优高斯随机游走和个体筛选策略的差分进化算法[J]. 控制与决策,

2016, 31(8): 1379-1386.

[6] 周丽燕,王立洪. 长记忆多项式回归模型的置信域估计[J]. 应用数学学报, 2014, 37(1): 31-44.

[7] 芮兰兰,李钦铭. 基于组合模型的短时交通流量预测算法[J]. 电子与信息学报, 2016, 38(5): 1227-1233.

[8] 李祖雄. 一类具有反馈控制的修正 Leslie-Gower 模型的周期解[J]. 应用数学学报, 2015, 38(1): 37-52.

[9] Din D R, Huang J S. Multicast backup reprovisioning problem for Hamiltonian cycle-based protection on WDM networks[J]. Optical Fiber Technology, 2014, 20(2): 142-157.

[10] Choi J S. Design and implementation of a PCE-based software-defined provisioning framework for carrier-grade MPLS-TP networks [J]. Photonic Network Communications, 2014, 29(1): 96-105.

作者简介:

单冬红 女(蒙古族), (1976-), 硕士研究生, 副教授. 研究方向为数据挖掘、计算机网络. E-mail: wind123456@189. cn.

吕琼帅 男, (1985-), 硕士研究生, 讲师. 研究方向为计算机网络、数据挖掘.

赵伟艇 男, (1967-), 硕士研究生, 教授. 研究方向为计算机网络.

(上接第 91 页)

供了一种全新的思路. 系统要求低、安装方便、使用简单、加密效果显著, 可广泛应用于政府、军队、企业等场所.

参考文献:

[1] 李秋华,何为,朱桂斌. 基于 RFID 的磁盘加密系统[J]. 信息网络安全, 2012, 12(9): 27-31.

[2] 刘冰. C++ 程序设计教程: 基于 Visual 2008[M]. 北京: 机械工业出版社, 2009.

[3] 谭文,陈铭聚. Windows 内核安全与驱动开发[M]. 北京: 电子工业出版社, 2015.

[4] 刘荣. 圈圈教你玩转 USB[M]. 北京: 北京航空航天大学出版社, 2009.

[5] John Hyde,孙耀国,赵德刚. USB 设计应用实例[M]. 北京: 中国铁道出版社, 2003.

[6] Jan Axelson. USB 开发大全[M]. 4 版. 北京: 人民邮电出版社, 2011.

[7] 赛普拉斯总部. AN61347-NX2LP-Flex USB to NAND Flash Firmware Design Notes [EB/OL]. [2016-05-08]. [http://www. cypress. com/d-ocumentation/appli-cation-notes/an61347-nx2lp-flex-usb-nand-flash-firm-ware-design-notes](http://www.cypress.com/d-ocumentation/application-notes/an61347-nx2lp-flex-usb-nand-flash-firmware-design-notes).

[8] 赛普拉斯总部. AN64408-Getting Started with NX2LP-Flex? [EB/OL]. [2016-03-08]. <http://www. cypress. com/documentation/Applicat-ion-notes/an64408-getting-started-nx2lp-flex>.

作者简介:

何 为 男, (1977-), 硕士, 讲师. 研究方向为模式识别、信息安全. E-mail: cqchhw@126. com.

于 洋 女, (1991-), 学士. 研究方向为信息安全、无线通信.

晁志超 男, (1982-), 博士, 讲师. 研究方向为图形处理、信息安全.