

基于 PSR-LSSVM 的网络流量预测

陈卫民¹ 陈志刚²

(湖南城市学院信息科学与工程学院 益阳 413000)¹ (中南大学软件学院 长沙 410000)²

摘要 为了提高网络流量预测精度,利用相空间重构和预测模型参数间的相互联系,提出一种遗传优化最小二乘支持向量机的网络流量预测方法。首先将相空间重构和最小二乘支持向量机参数作为遗传算法的个体,将模型预测精度作为个体适应度函数,然后通过遗传操作获得模型全局最优参数,最后通过网络流量仿真实验进行性能测试。结果表明,相对于传统预测方法,遗传优化最小二乘支持向量机提高了网络流量的预测精度,为网络流量预测提供了一种新的研究思路。

关键词 网络流量,相空间重构,最小二乘支持向量机,遗传算法

中图分类号 TP393 **文献标识码** A

Network Traffic Prediction Based on Phase Space Reconstruction and Least Square Support Vector Machine

CHEN Wei-min¹ CEHN Zhi-gang²

(School of Information Engineering and Science, Hunan City University, Yiyang 413000, China)¹

(School of Software, Central South University, Changsha 410000, China)²

Abstract In order to improve the network traffic prediction accuracy, this paper proposes a network traffic prediction method based on least square support vector machine(LSSVM) optimized by genetic algorithm which uses the relation between phase space reconstruction and parameters of prediction model. Firstly, phase space reconstruction and the parameters of LSSVM were used as an individual of genetic algorithm while the model prediction accuracy was used as the fitness function, and then global optimal parameters of the model were obtained by genetic algorithm, lastly, the simulation tests were carried out based on network traffic data. The results show that, compared with the traditional forecasting methods, the proposed model improves the prediction accuracy of network traffic and provide a new research thought for network traffic prediction.

Keywords Network traffic, Phase space reconstruction, LSSVM, GA

1 引言

随着网络规模和业务种类的不断增长,高速宽带网络管理的任务也越来越繁重,网络流量预测是网络管理的基础。一个预测精度高的模型不仅能够对流量进行预测,且可应用于网络带宽分配和网络拥塞控制^[1]。

针对网络流量预测问题,专家和学者进行了深入的研究,提出了许多有效的、切实可行的预测模型。传统网络流量预测模型有:线性回归分析、灰色模型和时间序列等^[2-4],这些方法都基于线性建模,不能准确地刻画网络流量的非平稳性、突变和混沌等特点,预测精度低^[5]。近年来出现的基于混沌理论的网络流量预测模型,其包括相空间重构和预测模型参数优化两个关键问题^[6]。相空间重构用于对隐藏于网络流量数据间的信息进行深入挖掘,对原动力系统描述;预测模型用于对网络流量数据进行学习和建模^[7]。最小二乘支持向量机(LSSVM)是基于结构风险最小化原则的算法,具有很强的自学习能力和逼近任意非线性系统的能力,其克服了传统机

器学习算法过拟合、局部最优的缺陷,泛化能力优异,成为网络流量的主要预测算法^[8],因此本研究选择 LSSVM 作为网络流量的预测模型。在基于的 LSSVM 网络流量预测过程中,相空间重构与 LSSVM 参数共同决定了网络流量的预测精度,两者互相联系,然而传统上两者常单独优化,忽略了两者间的内在联系,这样难以获得全局最优网络流量预测。

为了使网络流量预测模型性能达到最佳,相空间重构与 LSSVM 参数必须联合优化,为此,针对当前网络流量预测模型参数单独、分开优化存在的不足,提出了基于遗传优化最小二乘支持向量机的网络流量预测方法(GA-LSSVM)。采用遗传算法(Genetic Algorithm, GA)对相空间重构与 LSSVM 参数进行联合优化,并通过具体仿真实验对其性能进行验证。

2 相空间重构和最小二乘支持向量机

2.1 相空间重构

设观测到的网络流量时间序列为: $\{x(t)\}; t=1, 2, \dots, n$, 其中 n 表示样本数。根据 Takens 定理,通过确定最佳延迟时间

到稿日期:2011-09-20 返修日期:2011-12-15 本文受国家自然科学基金(61073186)资助。

陈卫民(1972-),男,硕士,讲师,主要研究方向为计算机网络和计算机应用, E-mail: 35069820@qq.com; 陈志刚(1962-),男,教授,博士生导师,主要研究方向为计算机网络、分布式系统。

(τ)和嵌入维(m)将网络流量数据重构为:

$$X_n = (x_n, x_n + \tau, \dots, x_n + (m-1)\tau) \quad (1)$$

从式(1)可知,相空间重构结果的优劣由 τ 和 m 决定,而相空间重构质量直接影响着模型的预测结果。根据 Takens 定理,没有噪声、无限的网络流量时间序列与 m 没有联系。然而,在实际应用中,网络流量数据长度不能无限长且含有噪声,因此,大多数学者认为 τ 和 m 应该联合优化。当前 τ 和 m 联合优化方法主要有时间窗口法、C-C 法(Coupled Cluster method)和自相关函数法等,这些方法性能各异,判断标准具有较强主观性^[9]。

2.2 LSSVM

最小二乘支持向量机(LSSVM)是 Suykens 等在标准支持向量机(SVM)的基础上提出的一种新的人工智能学习算法。LSSVM 将 SVM 的损失函数设定成误差平方和,以减少模型需要优化的参数,降低求解复杂性^[10]。

设获得网络流量数据集为: $\{(x_i, y_i)\}, i=1, 2, \dots, n$, 非线性函数 $\varphi(\cdot)$ 将数据集映射到高维特征空间进行优线性回归,即有:

$$f(x) = w^T \varphi(x) + b \quad (2)$$

式中, b 表示偏置量, w 表示权值向量。

根据 LSSVM 的原理,将式(2)问题求解转化成一个二次优化问题,即:

$$\begin{aligned} \min \quad & \|w\|^2 + \frac{1}{2} \gamma \sum_{i=1}^n \xi_i^2 \\ \text{s. t. } & y_i - w^T \varphi(x_i) + b = e_i, (i=1, 2, \dots, n) \end{aligned} \quad (3)$$

式中, γ 表示惩罚参数, e_i 为模型预测误差。

通过引入拉格朗日乘子,将式(3)转换为对偶优化问题,即:

$$\begin{aligned} L(w, b, \zeta, a) = \min \quad & \|w\|^2 + \frac{1}{2} \gamma \sum_{i=1}^n \xi_i^2 + \sum_{i=1}^n a_i (w^T \varphi(x_i) - \\ & b + e_i - y_i) \end{aligned} \quad (4)$$

式中, a_i 表示拉格朗日乘子。

对非线性回归问题,通过引入核函数将其转换成非线性预测,根据 Mercer 条件,核函数定义为:

$$\begin{aligned} K(x_i, x_j) &= \varphi(x_i)^T \varphi(x_j) \\ \text{径向基核函数定义为:} \\ k(x_i, x_j) &= \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \end{aligned} \quad (5)$$

式中, σ 表示径向基核数宽度。

本研究采用径向基核函数作为 LSSVM 的核函数, LSSVM 回归模型为:

$$f(x) = \sum_{i=1}^N a_i \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) + b \quad (6)$$

当前 LSSVM 参数 γ 和 σ 优化方法主要有遗传算法(GA)、粒子群优化算法(Particle Swarm Optimization, PSO)和网格搜索算法等^[11]。网格搜索算法速度慢,内存消耗多,运算量大,难以获得全局最优解,容易陷入局部最优;而传统的基于 GA 和 PSO 的算法没有考虑 LSSVM 参数和相空间重构间的联系,只对 LSSVM 参数进行优化,不能保证找到全局最优参数。

3 GA-LSSVM 的网络流量预测模型构建

τ 、 m 、 γ 和 σ 的联合优化是一个组合优化问题,传统穷举

法耗时长,本研究采用 GA 对模 τ 、 m 、 γ 和 σ 进行联合优化。

3.1 遗传算法

遗传算法(GA)是美国密歇根大学 Holland 教授提出的进化算法,用以模拟适者生存机制,其通过遗传机制对种群进行进化,找到问题的最优解。其工作流程如图 1 所示^[12]。

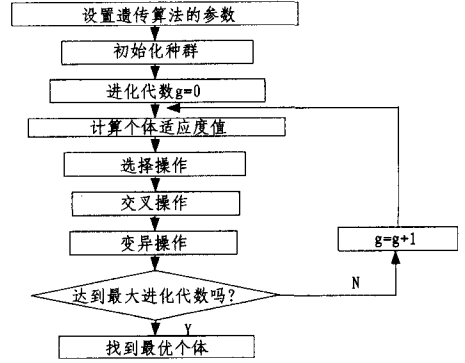


图1 遗传算法的工作流程

3.2 染色体编码设计

每一个染色体代表一个参数组合,本研究采用二进制编码方法对染色体进行编码,每一个染色体包括 4 部分: τ 、 m 、 γ 和 σ , 具体如图 2 所示。

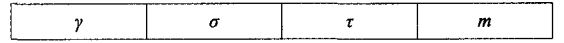


图2 染色体编码设计

模型实际参数通过式(7)转换得到:

$$p = \min p + \frac{\max p - \min p}{2^l - 1} \times d \quad (7)$$

式中, $\min p$ 和 $\max p$ 分为参数的最小值和最大值, l 为位串长度, d 为二进制位串。

3.3 适应度函数设计

适应度函数用来衡量一个个体的优劣,网络流量预测模型参数优化的目标是提高网络流量的预测精度,因此本研究将个体适应度函数 $f(x)$ 定义为:

$$f(x) = \frac{1}{\text{MSE}(\tau, m, w, \theta)} \quad (8)$$

式中, MSE(Mean Squared Error)表示均方误差,其定义如下:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (9)$$

式中, n 表示网络流量训练样本大小, y_i 和 $\hat{y}_i$ 分别表示网络流量期望值和预测值。

3.4 遗传操作

(1)选择操作。假定个体 x_i 的适应度为 $f(x_i)$, 则该个体被选择的概率 p_s 为:

$$p_s = \frac{f(x_i)}{\sum f(x_j)} \quad (10)$$

式中, $\sum f(x_j)$ 表示群体中的个体适应度的总和。

(2)交叉操作。在染色体的 4 段上分别选择 4 个交叉点进行交叉,交叉完后,如果其值不在参数取值范围内,则重新交叉。

(3)变异操作。根据变异概率选择一个染色体,然后随机选择一个变异点,采用 $0 \rightarrow 1, 1 \rightarrow 0$ 的规则进行变异,变异后,如果其值不在参数取值范围内,则重新变异。

3.5 网络流量预测模型工作步骤

Step1 随机生成规模 N_{pop} 的初始种群 R_{gen} 。

Step2 进化次数 $g=0$ 。

Step3 对染色体进行反编码,根据 τ 和 m 进行重相空间重构,LSSVM 根据 γ 和 σ 进行建模,得到每一组参数的网络流量预测精度,通过预测结果的 MSE 对个体适应度值进行计算。

Step4 如果达到算法的最大进化代数或者种群的最优个体所对应的适应度函数值没有发生大的变化,则表示参数优化结束,跳转到 Step7。

Step5 进化代数加 1,即 $g=g+1$ 。

Step6 根据选择、交叉和变异算子对种群进行更新,并跳转到 Step3。

Step7 得到网络流量预测模型的最优参数 τ 、 m 、 γ 和 σ 值。

Step8 采用 τ 、 m 对网络流量时间序列进行重构, γ 和 σ 作为 LSSVM 的参数进行训练建模,对测试集进行预测,输出模型对网络流量预测的结果。

4 仿真实验

4.1 数据集来源

实验数据来源于网络流量文库: <http://newsfeed.ntcu.net/~news/2011/>,收集了主节点路由器 Incoming articles 从 2011 年 6 月 1 日到 10 月 25 日每小时网络访问流量,得到 2000 个数据。为了尽量避免数据和计算中的偶然性造成评估结果的不准确,以前 1500 个数据作为训练集建立网络流量预测模型,后 500 个数据作为测试集进行预测检验,具体如图 3 所示。

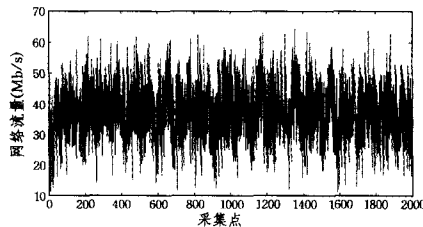


图 3 网络流量数据

4.2 网络流量数据归一化处理

由于 LSSVM 采用核函数,因此在进行训练时,处于 $[0,1]$ 的数据最为灵敏。从图 3 可知,实际网络流量具有突变性,变化幅度较大,为了提高 LSSVM 训练的效率,在建模之前对网络流量数据进行归一化处理,具体为:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (11)$$

最后对模型的网络流量预测结果进行反归一化处理,恢复真实预测值:

$$x = x'(x_{\max} - x_{\min}) + x_{\min} \quad (12)$$

式中, x 表示网络流量原始数据, $x_{\min}$ 和 $x_{\max}$ 分别表示最小值和最大值。

4.3 模型评价指标

为说明 GA-LSSVM 的有效性,在 Matlab 2007 平台编程实现算法。为了使 GA-LSSVM 预测结果具有可比性和说服力,采用 CC-LSSVM 作为对比模型(相空间重构采用 C-C 方

法,优化 LSSVM 参数采用 GA)。模型性能的评价指标为均方根误差(Root Mean Squared Predict error, RMSE)和平均绝对百分误差(Mean Absolute Percentage Error, MAPE),定义如下:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (13)$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{(y_t - \hat{y}_t)}{y_t} \right| \times 100 \quad (14)$$

式中, y_t 表示网络流量实际观测值, $\hat{y}_t$ 表示预测值。

4.4 结果与分析

4.4.1 CC-LSSVM 和 GA-LSSVM 的拟合结果比较

将训练样本分别输入到 GA-LSSVM 和 CC-LSSVM 进行学习,建立相应的网络流量预测模型,然后采用建立的模型对网络流量的训练样本进行拟合,得到的模型拟合结果如图 4 所示。

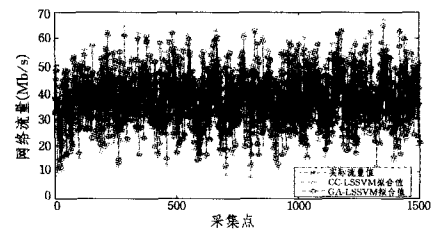


图 4 CC-LSSVM 和 GA-LSSVM 的拟合结果

从图 4 可知,相对于 CC-LSSVM,GA-LSSVM 网络流量值与网络流量实际值更加吻合,GA-LSSVM 拟合结果更好,拟合结果初步表明本研究提出的 GA-LSSVM 的参数联合优化思想是正确的、可行的。

4.4.2 CC-LSSVM 和 GA-LSSVM 的泛化能力比较

由于评价一个预测模型性能的优劣,主要考察其预测能力而非拟合能力,因此,分别采用 GA-LSSVM 和 CC-LSSVM 对测试集进行预测,其结果如图 5 所示。根据图 5 可知,相对于 CC-LSSVM,GA-LSSVM 预测精度更高,预测误差大大降低,是一种高精度的网络流量预测方法,可以很好地对网络流量变化趋势进行捕捉。

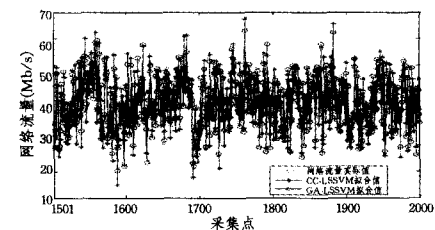


图 5 CC-LSSVM 和 GA-LSSVM 对测试集的预测结果

4.4.3 CC-LSSVM 和 GA-LSSVM 的整体性能比较

各种模型的预测拟合和预测误差如表 1 所列。从表 1 各种评价指标可知,GA-LSSVM 各项指标均优于 CC-LSSVM。对比结果表明,对相空间重构和 LSSVM 参数采用遗传算法进行联合优化,充分利用了两者的内在联系,可以提高网络流量预测模型的整合等性能,解决传统两者单独优化存在的缺陷;而 GA-LSSVM 能够更加精确刻画网络流量变化趋势和变化特点,预测结果更加可靠。这进一步说明对网络流量预测模型的参数进行联合优化是必需的。

表1 CC-LSSVM和GA-LSSVM性能对比

样本集	CC-LSSVM		GA-LSSVM	
	RMSE	MAPE	RMSE	MAPE
训练集	4.31	4.87%	0.17	0.97%
测试集	11.77	8.56%	0.60	1.93%

结束语 由于受到多种因素的综合影响,网络流量具有时变性、混沌性。针对该特点,本研究提出了一种 GA-LSSVM 的网络流量预测模型。结果表明,相对于传统的预测模型,GA-LSSVM 的网络流量预测结果误差更小,预测精度进一步提高,能更好地满足大规模网络流量预测的实时性需求。

参考文献

- [1] 谢高岗,闵应骅,张大方,等.一个基于实际测试的网络流量模型[J].计算机工程与科学,2001,23(5):51-54
- [2] 高姜明,吴春明,胡大民,等.网络流量预测中的时间序列模型比较研究[J].电子学报,2009,37(11):2353-235
- [3] 孙韩林,金跃辉,崔毅东,等.粗粒度网络流量的灰色模型预测[J].北京邮电大学学报,2010,33(1):7-11
- [4] 侯家利. Elman 神经网络的网络流量预测[J].计算机仿真,2011,28(7):154-156
- [5] Silva C G. Time series forecasting with a nonlinear model and

the scatter search meta-heuristic[J]. Information Sciences, 2008,178(16):3288-3299

- [6] Wang W S, Jin J J, Li Y Q. Prediction of inflow at three gorges dam in yangtze river with wavelet network model[J]. Water Resources Management, 2009,23(13):2791-2803
- [7] 陆锦军,王执铨.基于混沌特性的网络流量预测[J].南京航空航天大学学报,2006,38(2):217-221
- [8] 叶美盈,汪晓东,张浩然.基于在线最小二乘支持向量机回归的混沌时间序列预测[J].物理学报,2005,54(6):2568-2573
- [9] 马红光,李夕海,王国华,等.相空间重构中嵌入维和时间延迟的选择[J].西安交通大学学报,2004,38(4):335-338
- [10] 纪玲玲,林振山,王昌雨,等.最小二乘回归支持向量机对非线性时间序列预测的试验分析[J].解放军理工大学学报:自然科学版,2009,10(1):92-98
- [11] 吴景龙,杨淑霞,刘承水.基于遗传算法优化参数的支持向量机短期负荷预测方法[J].中南大学学报:自然科学版,2009,40(1):180-184
- [12] 陈果.基于遗传算法的支持向量机时间序列预测模型优化[J].仪器仪表学报,2006,27(9):1080-1084
- [13] 石贵民,林宏基.基于旁路的网络流量监控模式[J].重庆工学院学报:自然科学版,2011,25(9):63-69

(上接第 57 页)

10%~60%时两种三维算法定位精度的比较情况。从仿真结果可知,在节点总数不变的情况下,两种定位算法的定位精度总体上都随锚节点的增加而提高。从图上可以看出,在锚节点比例小于 20%时,APIS 定位算法优于本文算法,主要是因为本文的定位算法在锚节点较少时,在未知节点通信范围内无法获得更多的参考点进行准确估计,随着锚节点数的增加,利用本文的三维定位算法则可以估计出更精确的位置。从总体上看,本文算法的定位精度优于 APIS 定位算法。

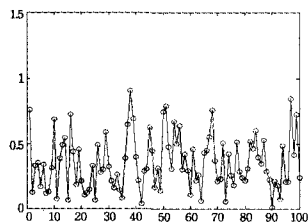


图7 每个未知节点的定位精度

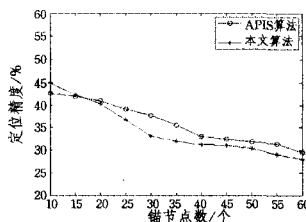


图8 锚节点比例对定位误差的影响

结束语 节点定位在无线传感器网络的应用中起着重要作用,尤其是三维定位有着更广阔的应用前景。本文提出了一种新的无线传感器网络三维定位算法,有效地解决了无线传感器网络的空间定位问题。仿真实验表明,本文的三维定位算法通信开销小,进一步减少了节点能量消耗,且具有较高的定位覆盖率,在网络规模一定的情况下,与 APIS 定位算法相比,定位精度优于 APIS 算法。本文算法仍需要布设一定数量的锚节点来计算校正值,虽然数量不大,但增加了成本。下一步将致力于研究采用单个移动锚节点来对三维空间中未知节点进行定位,使算法更加优化。

参考文献

- [1] Demirkol I, Ersoy C, Alagoz F. MAC protocols for wireless sensor networks; a survey [J]. IEEE Communications Magazine, 2006,44(4):115-121
- [2] 杜治高,钱德沛,刘铁.无线传感器网络中的地址分配协议[J].软件学报,2009,20(10):2787-2798
- [3] Cullar D, Estrin D, Strvstava M. Overview of sensor network [J]. Computer, 2004,37(8):41-49
- [4] 丘岩,赵冲冲,戴桂兰,等.无线传感器网络节点定位技术研究[J].计算机科学,2008,35(5):47-50,63
- [5] Zhang Li-qiang, Zhou Xiao-bo, Cheng Qiang. Landscape-3D: A Robust Localization Scheme for Sensor Networks over Complex 3D Terrains[C]// Proc. of 31st IEEE Conference, South Bend, IN; Indiana Univ., 2006:239-246
- [6] Liang Jian-lin, Shao Jun, Xu Ying, et al. Sensor Network Localization in Constrained 3-D Spaces[C]// Proc. of IEEE International Conference on Mechatronics and Automation, Luoyang, China, 2006:49-54
- [7] 吕良彬,曹阳,高洵,等.基于球壳交集的传感器网络三维定位算法[J].北京邮电大学学报,2006,29(5):48-51
- [8] Sundaramurthy M C, Chayapathy S N, Kumar A. Wi-Fi assistance to SUPL-based assisted-GPS simulators for indoor positioning[C]// IEEE Consumer Communications and Networking Conference(CCCN' 2011). San Antonio, TX, USA, 2011:918-922
- [9] Wang E, Zhang Shu-fang, Zhang Zhi-xian. Research on the Incomplete Constellation GPS Positioning Algorithm by a Combined Altitude and Clock Bias Model[C]// Information Engineering(ICIE). Shenyang, China, 2010:219-222
- [10] 查普曼. MATLAB 编程(第四版)[M]. 北京:科学出版社,2011
- [11] 张葛祥,李娜. MTLAB 仿真技术与应用[M]. 北京:清华大学出版社,2003