一种基于相关分析的局域最小二乘支持向量机小尺度网络流量预测算法*

唐舟进† 彭涛 王文博

(北京邮电大学信息与通信工程学院, 北京 100876)

(2014 年 1 月 15 日收到; 2014 年 4 月 11 日收到修改稿)

本文分析了网络流量数据的特性, 针对传统预测算法在预测网络流量时的缺陷提出了一种基于相关分析的相关局域最小二乘支持向量机 (LSSVM) 预测算法. 算法在对训练数据重构相空间后, 利用相关分析同时从距离相关和时间相关的训练样本中选择最优的训练子集, 结合自适应参数优化的 LSSVM 预测模型对小尺度网络流量进行预测. 通过选用实际情况下的网络流量数据对算法进行测试验证, 结果显示本文所提算法不仅优于传统的全局预测算法, 同时也优于各种改进的局域预测算法. 算法不仅在预测精度上取得大幅的性能提升, 同时能够通过留一交叉验证法在预测之前就完成预测模型和训练子集的优化.

关键词: 网络流量预测, 混沌时间序列预测, 最小二乘支持向量机, 局域预测

PACS: 05.45.Tp, 05.45.Gg **DOI:** 10.7498/aps.63.130504

1 引 言

目前计算机网络技术发展迅速, 通过网络提供的服务和应用日益丰富, 这也意味着对网络容量和利用率有着更高的要求. 加强网络管理技术的应用能够改善当前网络的利用率, 解决网络拥塞问题. 网络流量的预测技术能够在网络管理中起到关键性的作用, 准确的网络流量预测能够预防网络拥塞, 同时能够揭示网络的动态特性, 对网络设计和流量控制都有积极的意义.

时间序列预测技术可以按照预测时间长度的大小进行分类, 例如以天 (d) 为单位进行流量预测可以认为是大尺度的时间序列预测. 当前流量预测技术的研究主要集中于以大时间尺度或者中等时间尺度的流量预测方面 [1−13], 大时间尺度往往以 d 为预测单位 [1−3], 而中等时间尺度的流量预测则是以 h [4−6] 或者 min [7−13] 为预测单位. 更小的时间尺度, 如以秒为预测时间单位的流量预测技术研究较少. 小尺度网络流量序列的特性与较大尺度的网流量序列是不同的, 直接套用现有的预测技术对小尺度网络流量进行预测不可靠. 这是因为大尺度的流量预测技术依赖的训练数据采样时间间隔较大, 反映的主要是网络流量的趋势, 序列变化平滑, 同时训练数据集的规模也较大, 预测的难度较低.

而在小尺度的网络流量预测方面, 由于网络承载的业务种类繁多, 网络流量的变化趋势在小尺度的时间范围内表现出的变化特性十分复杂, 具有突发性强的特点, 准确预测的难度较大. 但预测时间尺度较小也意味着系统能够对网络流量的变化进行即时的反应, 能够通过对变化趋势的把握进行流量管控, 有利于解决网络拥塞的问题.

但小尺度网络流量预测技术的应用还需要解决以下几个问题:

1) 由于网络用户行为的复杂性和变化性, 网络环境的变化也较为剧烈. 因此预测技术需要自动适应网络环境的变化, 在各种环境下都能够保持较高的预测精度, 并且这种自适应必须是实时的, 不需要人工参与.

2) 网络流量预测技术必须有能力识别流量数

* 国防科技预研项目 (批准号: 208010201) 资助的课题.

† 通讯作者. E-mail: tangzhoujin@gmail.com

© 2014 中国物理学会 Chinese Physical Society

据序列中的异常数据. 这些异常数据往往是由于网络的突发性造成的, 具有噪声的特性, 会降低预测的精度.

3) 大时间尺度下的长相关性和小时间尺度下的自相似性 (分形性) 是网络流量公认的最重要的统计特征. 这也意味着网络流量具有非线性和混沌特性, 预测算法需要对非线性变化具有良好的预测能力.

目前, 非线性时间序列预测算法得到了广泛的研究[5,10−16], 但是这些算法在预测网络流量时都有自身的缺陷. 例如, 在流量预测中应用最为广泛的非线性预测算法主要是基于神经网络[13,14,16]. 但基于神经网络的算法有几个缺陷: 1) 神经网络预测模型的结构还没有一种有效的设计方法, 主要依赖经验进行设计; 2) 神经网络算法的收敛速度较慢, 难以满足流量预测实时性的要求; 3) 神经网络容易出现过拟合的现象.

支持向量机 (SVM) 预测模型引入了结构风险最小化原则和统计学习理论的研究成果[17], 克服了神经网络的缺陷. 同时支持向量机有多种改进算法, 其中 LSSVM 算法[18] 将 SVM 算法中的不等式约束转换为等式约束, 降低了算法的复杂度, 在混沌时间序列预测中得到了较好的应用效果[19]; 文献 [20] 提出的多维循环 LSSVM, 在传统 LSSVM 算法基础上通过将训练集循环映射到高维相空间提高了混沌时间序列的预测精度; 文献 [21] 引入模糊 sigmoid 核函数替换传统 SVM 算法中常用的径向基 (RBF) 核函数, 通过模糊逻辑方法大幅度降低了算法运行时间.

尽管有不少优点, 但传统的 LSSVM 预测算法仍然难以直接应用于网络流量的预测. 首先, 因为传统 LSSVM 算法是一种离线预测的算法, 通过训练集建立预测模型之后, 其参数和训练样本不再发生变化. 而网络流量是时变的, 其突发性较强. 离线预测适用于动力系统时不变或者慢变的混沌时间序列预测, 对网络流量序列则应当采用在线预测的方式, 提高其自适应的能力. 其次, 由于网络流量具有长相关性, 如果通过全部训练数据进行训练, 需要较长时间内的采样数据进行训练, 而 LSSVM 预测算法的复杂度高达 $O(N^3)$, 其中 N 是训练数据的长度. 这就无法满足实时性的要求.

本文针对上述问题, 提出了一种基于相关分析的相关局域 LSSVM 小尺度网络流量预测算法. 首先, 算法利用网络流量的自相似性和长相关性, 同时从距离相关和时间相关两方面对训练数据进行筛选. 然后算法通过参数组合的联合寻优, 优化 LSSVM 预测模型参数, 同时对筛选出的训练子集进行进一步的优化. 最后, 算法的优化是通过留一交叉验证法完成, 能够仅依靠训练数据就完成参数和训练子集的优化.

2 小尺度网络流量混沌特性分析

混沌是指在确定性系统中出现的一种类似无规则的、类似随机的现象. 如何判断混沌系统的特性, 仍是混沌工程学的重要研究课题. 一般地, 判断系统混沌特性的方法有: 相轨线图分析法, 自功率谱分析法, 李雅普诺夫指数法等. 本文选用相轨线图分析法和李雅普诺夫指数法对小尺度网络流量数据进行分析.

实际的小尺度网络流量数据序列来源于 Lawrence Berkeley National Laboratory 发布的 TCP 流量数据 (http://ita.ee.lbl.gov/), 该数据为 1 h 内实验室服务器与 Internet 网络进行交互的流量数据. 数据的具体细节可以参见实验室的网页. 数据包选用 DEC-Pkt1, 原始采样精度为毫秒, 为了利于试验测试对比, 本文对原始数据进行间隔为 0.1 s 的重采样, 得到长度为 36000 的流量序列, 每次采样数据为 0.1 s 内到达的 TCP 包的总比特数. 流量数据如图 1 (a) 所示.

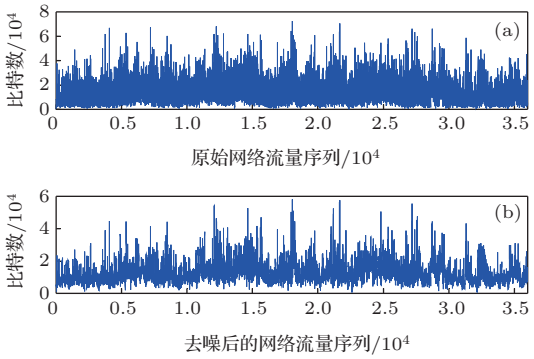

图 1 (a) 原始网络流量序列; (b) 去噪后的网络流量序列

首先采用李雅普诺夫指数法对流量数据的混沌特性进行分析, 李雅普诺夫指数表示相邻轨线间的平均发散率, 根据非线性动力学相关原理可知, 当某一序列的最大李雅普诺夫指数为正时, 可以判定该段序列存在混沌现象. 取延迟为 1, 嵌入维数为 2, 可得到原始数据的最大李雅普诺夫指数为 0.6879. 可以判定小尺度网络流量序列具有混沌特性.

然后用相轨线图分析法对序列的混沌特性进行分析, 序列的二相图如图 2 (a) 所示. 从图中可以看出, 序列的相轨线十分紊乱, 难以观察到混沌系统的无穷层次自相似结构, 随机性较强, 而有序性较弱. 主要是因为网络业务的突发性和复杂性使得采样数据中含有大量的噪声, 随机性远大于有序性. 本文采用文献 [16] 中提出的局域支持向量机预测算法对未去噪的原始流量序列进行预测, 发现预测结果存在一步延迟的情况, 即预测结果同实际流量值之间存在一个单位的滞后, 说明原始数据难以进行有效的预测.

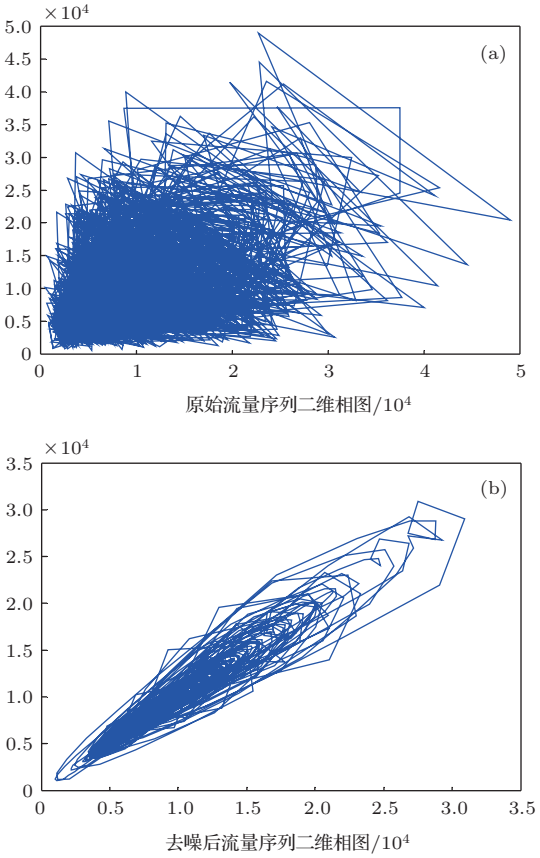

图 2 (a) 原始流量序列二维相图; (b) 去噪后流量序列二维相图

为了降低序列中的噪声成分, 同时为了方便预测算法性能对比, 本文选用小波软阈值法对原始数据进行降噪处理, 小波软阈值去噪方法具有降低分析信号的随机噪声成份的优良性能, 在小尺度网络流量预测中已经得到了应用 [14−16]. 本文采用 Matlab 小波工具箱中的一维小波去噪函数 wden 函数实现原始流量序列的小波软阈值去噪, 选择小波基为 bior3.3, 分解层数设置为 3, 阈值取值方式选择为 sqtwolog 固定阈值形式, 阈值处理形式设置为 s 软门限阈值形式, 噪声估计选择 sln 形式, 即只根据第一层小波系数估计噪声水平.

去噪后的网络流量序列如图 1 (b) 所示, 可以看出噪声明显减少. 对去噪后的流量序列同样进行混沌特性分析, 先计算序列的最大李雅普诺夫指数, 取延迟为 1, 嵌入维数为 2, 用 wolf 法得到去噪后序列数据的最大李雅普诺夫指数为 0.2092.

因此去除噪声后的网络流量趋势序列仍然具有混沌特性. 然后通过相轨线图分析序列的混沌特性, 去噪流量序列的二相图如图 2 (b) 所示, 可以看到, 去除噪声后, 流量序列表现出明显的嵌套自相似结构, 有序性得到了增强, 表现出明显的自相似性.

3 LSSVM 预测模型

由于 LSSVM 预测模型是一种较为成熟的技术, 并且本文对预测模型本身的原理和实现不做改进, 所以本节仅对 LSSVM 预测模型原理进行简单介绍, 主要引用本文作者在文献 [22] 中的 LSSVM 模型相关的介绍内容, 更详细的 LSSVM 模型的原理可以参见文献 [18]. LSSVM 是基于 SVM 改进后的一种算法, 通过引入等式化约束和最小二乘损失函数的方法, 使最优化问题的求解变为求解线性方程, 避免了解二次规划问题, 使得算法的复杂度降低, 相比 SVM, LSSVM 的运算速度较快. 基于 LSSVM 的回归预测可以描述为:

首先对训练数据 X 重构后, 获得相空间矩阵 $S=\{\boldsymbol{x}_i,y_i\}_{i=1}^{k}$, $\boldsymbol{x}_i$ 为 $m-1$ 维输入向量, y_i 为一维输出向量, k 为重构后的训练样本个数. $\boldsymbol{x}_i$ 与 y_i 之间通常为非线性关系, 因此将 $\boldsymbol{x}_i$ 映射到高维特征空间中, LSSVM 主要是在高维空间中对训练样本进行回归:

$$y=\boldsymbol{\omega}^{\boldsymbol{T}}\varphi(\boldsymbol{x})+b, \tag{1}$$

其中, $\varphi(\boldsymbol{x})$ 为非线性映射函数, $\boldsymbol{\omega}$ 为法向量, b 为偏置量, 对以上问题的求解可描述如下:

$$\begin{aligned}&\min_{\boldsymbol{\omega},b,e} \quad J(\boldsymbol{\omega},b,e)=\frac{1}{2}||\boldsymbol{\omega}||^2+\frac{C}{2}\sum_{i=1}^{k}\xi_i^2,\\ &\text{s.t.} \quad y_i-\xi_i=\boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+b, \quad i=1,\cdots,k,\end{aligned} \tag{2}$$

C 为惩罚因子, ξ_i 为训练误差. 为求解此优化问题, 可引入 Lagrange 函数

$$\begin{aligned}&L(\boldsymbol{\omega},b,e,\alpha)\\ &=J(\boldsymbol{\omega},b,e)+\sum_{i=1}^{k}\alpha_i\left[y_i-\xi_i-\boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)-b\right],\end{aligned} \tag{3}$$

其中 α_i, $i=1,\cdots,k$ 为Lagrange乘子, 由KKT条件得到如下关系式:

$$\begin{aligned}
&\frac{\partial L}{\partial \boldsymbol{\omega}} = 0 \to \boldsymbol{\omega} = \sum_{i=1}^{k} \alpha_i \varphi(\boldsymbol{x}_i),\\
&\frac{\partial L}{\partial \alpha_i} = 0 \to y_i - \xi_i - \boldsymbol{\omega}^{\mathrm{T}} \varphi(\boldsymbol{x}_i) - b = 0,\\
&\frac{\partial L}{\partial b} = 0 \to \sum_{i=1}^{k} \alpha_i = 0,\\
&\frac{\partial L}{\partial \xi_i} = 0 \to \alpha_i = C\xi_i. \qquad (4)
\end{aligned}$$

关系式的求解可化为

$$\begin{bmatrix} 0 & \boldsymbol{e}^{\mathrm{T}} \\ \boldsymbol{e} & Q + C^{-1}I \end{bmatrix} \times \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix}, \qquad (5)$$

其中 Q 是元素为 K_{ij} 的 $k \times k$ 阶核矩阵,

$$K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \varphi(\boldsymbol{x}_i), \varphi(\boldsymbol{x}_j) \rangle,$$

I 为单位矩阵, 向量 $\boldsymbol{e} = [1,\cdots,1]^{\mathrm{T}}$, 向量 $\boldsymbol{\alpha} = [\alpha_1,\cdots,\alpha_k]^{\mathrm{T}}$, 向量 $\boldsymbol{y} = [y_1,\cdots,y_k]^{\mathrm{T}}$.

定义 $Q_n = Q + \dfrac{I}{C}$, 就可以得到 $\boldsymbol{\alpha}$ 和 b 的表达式:

$$\begin{aligned}
b &= \frac{\boldsymbol{e}^{\mathrm{T}} Q_n^{-1} \boldsymbol{y}}{\boldsymbol{e}^{\mathrm{T}} Q_n^{-1} \boldsymbol{e}},\\
\boldsymbol{\alpha} &= Q_n^{-1}(\boldsymbol{y} - \boldsymbol{e} \times b). \qquad (6)
\end{aligned}$$

通过 (5) 式和 (6) 式可得到LSSVM的混沌时间序列回归模型为

$$\begin{aligned}
f(\boldsymbol{x}) &= \sum_{i=1}^{k} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b,\\
y &= f(\boldsymbol{x}). \qquad (7)
\end{aligned}$$

对应的预测输入样本为 $\boldsymbol{x}_{\mathrm{p}}$, 则得到的预测值为

$$y_{\mathrm{p}} = f(\boldsymbol{x}_{\mathrm{p}}) = \sum_{i=1}^{k} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_{\mathrm{p}}) + b. \qquad (8)$$

本文采用径向基核函数 (RBF) 作为核函数:

$$K(\boldsymbol{x}_{\mathrm{i}}, \boldsymbol{x}_{\mathrm{p}}) = \exp\left(-\left\|\boldsymbol{x}_{\mathrm{i}} - \boldsymbol{x}_{\mathrm{p}}\right\|^2 / 2\delta^2\right), \qquad (9)$$

式中 δ^2 是核函数的方差.

4 相关局域预测算法

从前文网络流量特性分析中可以看到, 小尺度网络流量的数据量较大. 如果要满足网络流量实时性和自适应的需求, 利用传统的非线性预测算法是难以实现的. 首先良好的网络流量预测算法不能过于复杂.

非线性预测可以分为全局预测法和局域预测法 [23], 全局预测法利用全部的历史数据来建立预测模型, 认为全部的数据都具有同等重要的信息含量, 用全部数据去拟合系统动力方程. 传统的神经网络, 支持向量机等预测算法均属于全局预测法, 它们通过全部的历史数据重构相空间, 建立模型输入输出对, 拟合系统方程, 完成预测值的输出. 通常全局预测算法由于计算量较大, 一般用于离线预测, 即建立好预测模型之后就不会重新估计模型参数. 而局域预测法利用部分历史数据, 通常是最邻近预测样本的历史数据来建立预测模型, 由于训练样本较少, 计算量相对也较小. 在文献 [16] 中, 作者提出了一种局域相关向量机的预测算法用于小尺度网络流量的预测, 取得了较好的效果. 其首要步骤就是筛选训练集, 通过选择时间上最邻近的部分历史数据作为训练子集, 降低了预测的复杂度, 能够实现在线预测, 从而适应小尺度网络流量的快速变化特性.

从前文中的分析可以知道, 网络流量具有长相关性, 即较早的历史数据中仍然与预测样本之间存在相关性. 从相图的角度来看, 就是由于无穷嵌套的自相似结构, 预测样本所在轨迹与最邻近轨迹的相似性是较强的, 而最邻近轨迹往往包含预测样本欧氏距离最近的样本. 因此欧氏距离较近的样本通常在预测样本点邻近轨迹上, 邻近轨迹由于混沌系统的分形特性是与当前轨迹相似的, 距离邻近样本与预测样本相关性较强, 通过距离邻近样本训练预测模型有利于提高预测精度. 选择训练子集时, 不应该排除历史数据中与当前预测样本距离邻近的数据作为训练子集. 而一般的局域预测算法只考虑到时间上邻近的相点, 在网络流量这种时变系统中时间相关性是存在的, 但距离相关性也需要考虑.

当然同时考虑时间和距离相关又面临新的问题, 在训练子集样本数有约束的条件下, 如何将距离相关的子集和时间相关的子集统一起来. 本文对此进行分析. 首先, 通过距离相关筛选训练子集的思想已经得到了应用, 例如文献 [24] 提出的基于自组织映射 (SOM) 的分类预测, 通过将训练数据进行聚类形成多个训练子集, 建立对应的预测模型, 通过欧氏距离判断当前预测样本所属的预测模型. 这是一种基于距离相关的离线预测算法, 在预测电力负荷的变化趋势时取得了较好的效果. 但离线预测难以适应小尺度网络流量的快变特性, 如果将文

献 [24] 的方法用于在线预测, 那么每一次预测都要对海量的历史数据进行聚类建模, 这就会使得算法复杂度大增. 后文对时间相关局域预测和距离相关局域预测在小尺度网络流量预测中的应用情况进行试验分析.

为了充分利用网络流量序列的长相关性和自相似性, 本文提出一种同时利用距离相关和时间相关的局域LSSVM 预测算法. 其步骤如下:

步骤 1 如果是算法初始, 则输入去噪后的历史流量数据 X, 否则用预测实际值更新并输入历史数据.

步骤 2 将 X 按照嵌入维数 m 和时间延迟 τ 重构相空间, 可得训练矩阵

$$S=\{\boldsymbol{x}_i, X(i+m-1)\}_{i=1}^{n-m+1}$$

和预测样本输入向量 $\boldsymbol{x}_{\rm p}$, n 是 X 的序列长度, m 是嵌入维数.

步骤 3 选择

$$T_{\rm train}=\{\boldsymbol{x}_i, X(i+m-1)\}_{i=n-m-k_1+1}^{n-m+1},$$

即时间上最邻近的 k_1 个向量作为时间相关训练子集.

步骤 4 选择 $D_{\rm train}$ 为

$$\{\boldsymbol{x}_i, X(i+m-1)\}_{i=1}^{n-m-k_1}$$

中 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_{\rm p}$ 距离最近的 k_2 个样本构成的距离相关子集, 可以采用滑动过滤窗的方法实现.

1) 将 Win 设置为长度为 k_2 的滑动窗, $D_{\rm train}$ 的初值为 $\{\boldsymbol{x}_i, X(i+m-1)\}_{i=1}^{k_2}$, Win 的初值为 $\{w_i\}_{i=1}^{k_2}$, 其中 w_i 为 $\boldsymbol{x}_{\rm p}$ 与 $\boldsymbol{x}_i$ 之间的欧氏距离.

2) Win 从 $\boldsymbol{x}_{k_2+1}$ 开始滑动, 如果 $\boldsymbol{x}_{\rm p}$ 与 $\boldsymbol{x}_{k_2+1}$ 之间的欧氏距离小于 Win 中的最大值, 即 $\mathrm{dis}(\boldsymbol{x}_{\rm p}, \boldsymbol{x}_{k_2+1}) < \max(\mathrm{Win})$, 则去掉 $D_{\rm train}$ 对应 $\max(\mathrm{Win})$ 的一行, 并将 $\boldsymbol{x}_{k_2+1}, X(k_2+m)$ 添加至 $D_{\rm train}$ 最后一行, 同时更新 Win, 去掉 $\max(\mathrm{Win})$, 在 Win 的末尾添加 $\mathrm{dis}(\boldsymbol{x}, \boldsymbol{x}_{k_2+1})$, 否则 $D_{\rm train}$ 和 Win 不变.

3) 窗口 Win 向前滑行过滤, 一直到 $\boldsymbol{x}_{n-m-k_1}$ 结束.

步骤 5 合并 $T_{\rm train}$ 和 $D_{\rm train}$ 为完整的训练子集 $S_{\rm train}$, 然后通过训练时得到的优化参数组合建立 LSSVM 预测模型, 对预测样本进行预测, 输出单步预测值.

步骤 6 预测测试集如果测试完, 则算法完成, 否则继续执行步骤 1 到步骤 5.

算法流程图如图 3 所示.

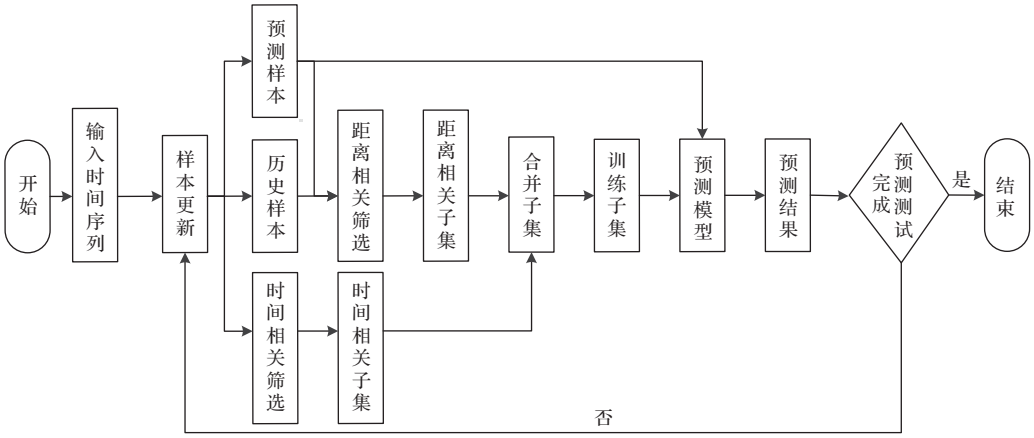

图 3 相关局域 LSSVM 预测算法流程图

5 算法参数优化

前文所提的算法还需要解决参数选择的问题. LSSVM 预测模型中的嵌入维数 m, 时间延迟 τ, 惩罚因子 C, 以及径向基核函数方差 δ^2 这些参数都会影响到预测精度. 传统的 LSSVM 算法将这些参数认为是独立的, 然后分别进行选择. 本文根据文献 [22] 的研究, 将参数组合 $[m, \tau, C, \delta^2]$ 进行联合寻优. 此外距离相关训练子集样本个数 k_2 和时间相关训练子集样本个数 k_1 也会影响预测精度, 由于预测的黑盒特性, 无法确定距离相关子集和时间相关子集的作用, 从而也无法确定子集规模和两种相关子集的比例关系. 但可以根据实时性的要求结合设备运算性能确定训练子集总的样本大小 k_3, 那么 k_1 和 k_2 的关系可以认为是 $k_3=k_1+k_2$. 其中 k_1, k_2 属于自然数, k_3 属于正整数. 那么在子集数为 k_3 这

个约束条件下, 只要确定 k_1 和 k_2 任意之一就能确定训练子集的选择. 为了使模式搜索的主要参数搜索上下界一致, 本文设置一个参数 k, k 与 k_1 的关系为 $k_1 = \lceil k/1000 \rceil$, 其中 $\lceil\rceil$ 表示向上取整. 将参数 k 也作为一个寻优参数加入参数组合中.

X 为历史数据, 通过留一交叉验证法[25](leave-one-out cross validation) 选择验证子集

$$\boldsymbol{y}_v = [X(n-s_v+1)\cdots, X(n-1), X(n)],$$

s_v 是验证子集的长度. 其余部分作为训练数据候选集, 通过前文中的相关局域LSSVM回归预测算法, 通过距离相关和时间相关筛选训练数据子集建立预测模型. 其中验证样本预测模型最优参数组合 $[m, \tau, C, \delta^2, k]$ 通过求解下式:

$$\min_{m,\tau,C,\delta,k} \left\{ \frac{1}{(s_v+1)} \sum_{i=1}^{s_v} \left[f'(m,\tau,C,\delta^2,k, \boldsymbol{x}_{n-m-s_v+i}) - \boldsymbol{y}_v(i) \right]^2 \right\}^{1/2},$$

$$f'\left(m,\tau,C,\delta^2,k,\boldsymbol{x}_{n-m-s_v+i}\right) = \sum_{j=1}^{k_3} \alpha'_j K\left(\boldsymbol{x}^{j'}_{T_i}, \boldsymbol{x}_{n-m-s_v+i}\right) + b'. \tag{10}$$

(10) 式中的 s_v 是验证子集的长度, $\boldsymbol{x}'_{T_i}$ 是算法根据参数 k 和验证输入样本 $\boldsymbol{x}^{n-m-s_v+i}$ 筛选出的训练子集组成样本, f' 是验证样本的LSSVM预测模型. 由于LSSVM 的黑盒特性, 难以直接求解最优参数组合. 尽管模式搜索算法在进行黑盒函数寻优时能否保证迭代过程的全局收敛性一直存在疑问[26], 但是模式搜索算法在许多工程问题中应用很好, 如文献[27]中, 作者利用多个数据库比较了模式搜索, 遗传算法, 网格搜索算法和试验设计法在选择LSSVM模型参数时的寻优能力, 发现模式搜索算法准确度最高, 收敛速度较快的特点. 所以本文选择模式搜索算法对LSSVM预测模型参数组合 $[m, \tau, C, \delta^2, k]$ 进行寻优.

由于时间上的相关性, 通过验证样本搜索到的模型参数对于预测样本也是较优的选择, 由验证样本的优选参数组合 $[m, \tau, C, \delta^2]$ 和训练子集参数 k_1 和 k_2 建立测试集的LSSVM预测模型, 由下式表达:

$$f\left(m,\tau,C,\delta^2,k,\boldsymbol{x}_{\mathrm{p}}\right) = \sum_{j=1}^{k_3} \alpha_j K(\boldsymbol{x}^j_T, \boldsymbol{x}_{\mathrm{p}}) + b. \tag{11}$$

值得注意的是, 本文提供的通过留一交叉验证和模式搜索法进行参数优化的方案, 由于验证子集本身就与预测测试集存在差异, 所以其搜索结果不是预测模型参数的全局最优解. 可以通过求解下式得到LSSVM预测模型的全局最优解, 但是这样就需要预测实际值进行参数优化, 无法在预测之前实现.

$$\min_{m,\tau,C,\delta,k} \sqrt{\frac{\sum_{i=1}^{s} \left[f\left(m,\tau,C,\delta^2,k,\boldsymbol{x}^i_{\mathrm{p}}\right) - \boldsymbol{y}(i) \right]^2}{(s+1)}}. \tag{12}$$

(12) 式中的 $\boldsymbol{y}$ 是预测测试集实际值, s 是预测测试集长度, 求解 (12) 式表示寻找与预测测试集误差最小的预测模型参数.

6 实验仿真与分析

如前文所述, 本文进行实验的数据选择为DEC-Pkt1流量包经过0.1 s重采样和去噪后得到的流量序列数据, 数据长度为36000. 其中前33000点作为历史数据, 后3000点作为预测测试集.

首先测试本文所提算法的预测性能, 为了方便对比, 本文选择正则化均方误差[14−16]作为预测性能评价指标, 其分母是历史数据的方差而非通常采用的测试集的方差.

$$\mathrm{NMSE} = \frac{\frac{1}{s}\sum_{i=1}^{s} |\boldsymbol{y}(i) - \boldsymbol{y}_{\mathrm{p}}(i)|^2}{\frac{1}{n}\sum_{i=1}^{n} \left[X(i) - \bar{X}\right]^2}, \tag{13}$$

式中 $n = 33000$, $s = 3000$, $\boldsymbol{y}_{\mathrm{p}}(i)$ 是模型预测输出值, $\boldsymbol{y}(i)$ 是预测集实际值, $\bar{X}$ 是历史数据均值.

本文通过留一交叉验证法, 选择

$$[x_{32000}, x_{32001}, x_{32002}, \cdots, x_{33000}]$$

作为验证样本, 通过模式搜索对本文所提的相关局域LSSVM预测算的参数组合进行优化. 在子集样本数为100的条件下, 优化参数组合为: 嵌入维数 $m = 7$, 时间延迟 $\tau = 1$, 惩罚因子 $\lambda = 4008$, 径向基核函数方差 $\delta^2 = 40.1$, 时间相关子集数 $k_1 = 13$, 距离相关子集数 $k_2 = 87$. 模式搜索算法的参数配置参见文献[22], 详细步骤参见文献[27].

将上述参数代入相关局域LSSVM预测模型中, 得到的预测结果如图4所示.

其中图4 (a) 是网络流量序列的实际值和预测值. 图4 (b) 是预测误差.

单步预测值的相对百分误差 (relative percent error, RPE) 如图4所示.

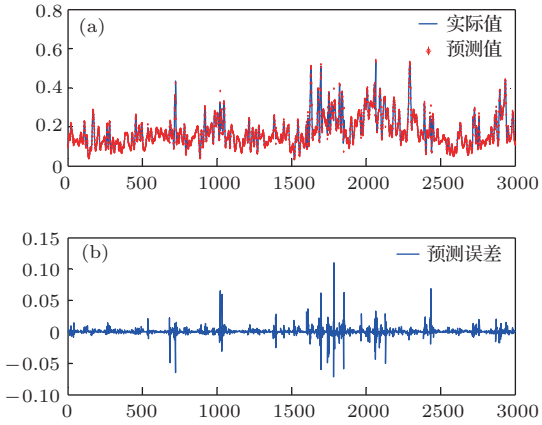

图 4 (a) 本文算法预测效果; (b) 算法预测误差

在图 4 (a) 中, 实线表示原始数据, 星号表示预测值, 可以看出, 预测值与实际值拟合较好. 说明本文所提算法适合用于小尺度网络流量序列的预测, 预测值的正则化均方误差仅为 3.531×10^{-3}. 从图 5 中可以看出预测值 RPE 的分布情况, 预测准确度超过 97.5% 的预测值所占比例为 87%, 预测精度超过 95% 的预测值所占比例为 95.6%, 说明绝大部分情况下, 算法都能够提供较为可靠的预测结果. 求解 (12) 式, 通过模式搜索法可以得到的预测模型参数组合最优解为嵌入维数 $m=7$, 时间延迟 $\tau=1$, 惩罚因子 $\lambda=2998$, 径向基核函数方差 $\delta^2=37$, 预测值的正则化均方误差为 3.372×10^{-3}. 可见本文算法预测精度相比理想情况下的最优预测模型性能下降不到 5%, 但却能在预测前完成模型参数的优化, 不需要预测实际值的优化, 具备了实用性.

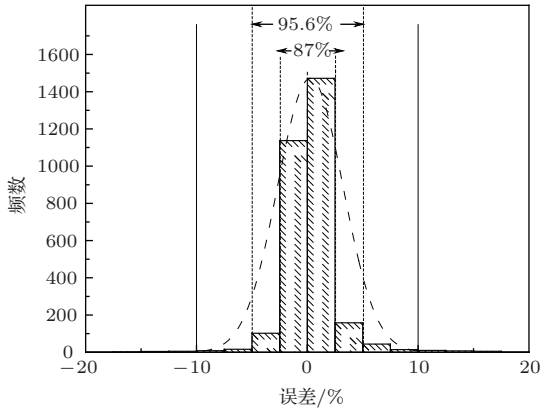

图 5 预测相对百分误差分布图

距离相关和时间相关的训练子集筛选方式对预测精度的影响各不相同, 本文通过设计 3 种不同类型的算法进行子集筛选方式影响的对比.

首先是只考虑时间相关的情况, 实验的算法 1 是根据文献 [15] 的局域 SVM 算法稍作改进将预测模型变为性能接近的 LSSVM 所设计的局域 LSSVM, 其训练子集仅由在时间上与预测样本最邻近的 k_3' 个历史数据样本组成.

然后是只考虑距离相关的情况, 本文对文献 [24] 所提出的分类预测算法稍作改进, 对每个预测样本点都建立一个预测模型, 通过全部历史数据中相对预测样本点欧氏距离最近的 k_3'' 个相点建立预测模型来实现在线预测, 这就是实验采用的分类 LSSVM 预测算法. 分类 LSSVM 可以看做一种较为极端的分类预测算法, 对于每一个训练样本都建立了一个预测模型.

最后是同时考虑距离相关和时间相关的情况, 即本文所提出的相关局域 LSSVM 预测算法, 其中 k_1, k_2 分别表示子集中时间相关样本和距离相关样本的数量, 通过模式搜索寻优确定, $k_3=k_1+k_2$. 实验时, 3 种算法仅训练子集不同, 其余参数均为前文所述的优化参数. 在训练子集样本数取值不同的条件下, 3 种算法的预测性能如图 6 所示.

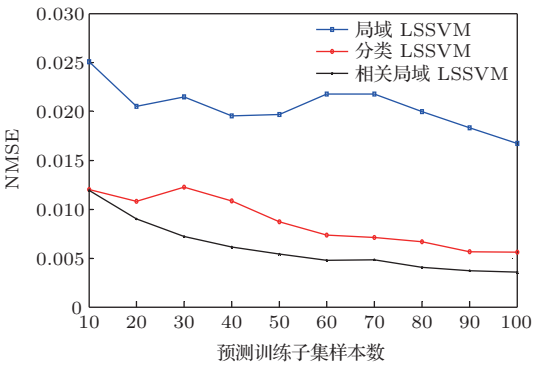

图 6 不同子集筛选方式对预测精度的影响

首先可以看到仅考虑时间相关的局域 LSSVM 预测性能较差, 这说明传统的局域预测算法并没有利用到网络流量序列的自相似特性, 从而使得预测精度在三种算法中最差, 当然其优点在于直接根据时间距离筛选训练子集, 不需要以前的历史数据进行运算, 运算量较小.

然后是只考虑距离相关的分类 LSSVM 算法, 从预测性能上看, 分类 LSSVM 算法的预测精度较高, 却显然忽略了网络流量序列具有时间相关的特性.

本文所提的算法预测精度最好, 复杂度和分类 LSSVM 相近, 相对局域 LSSVM 较高, 但比起传统的全局预测算法复杂度仍然较低. 从实验结果可以

看出, 网络流量具有长相关性和自相似性, 其时变性较强使得自相关衰减较快, 不利于传统局域预测算法的应用, 即使是较大的训练子集也无法带来较高的预测精度增益.

同时, 三种算法的预测精度均随着训练子集样本数的增大而提高, 精度提高的增幅逐渐减小. 这说明可以通过更大的训练子集提高预测精度, 但增加了算法的复杂度, 并且复杂度的增幅并非是线性增长, 精度越高, 那么同样的增幅需要增加的复杂度越大. 另外过大的训练子集容易使预测算法陷入过拟合的现象, 反而容易降低预测精度. 因此设计预测模型时, 需要根据实际的需求, 综合考虑预测精度与复杂度的关系.

表1是3种算法一次单步预测的复杂度和仿真耗时对比, 仿真平台配置的CPU为Pentium IV 2.5 GHz, 内存为2 GB RAM.

从表1中可以看出, 传统LSSVM运算量超过万亿, 过于复杂, 难以实现. 传统局域算法速度最快, 这是因为传统局域算法在子集筛选上几乎不消耗时间, 直接选取时间距离最近的若干样本进行预测, 并且由于网络流量的时间相关性衰减较快, 传统局域算法所需的样本数也较少. 距离相关的子集筛选需要消耗额外的运算量, 并且与维数和训练数据长度成正比, 这是因为在预测之前要遍历全部历史数据来筛选距离相关子集, 当历史数据较多时, 运算量就大于传统局域算法. 此外从前文的分析中可以看出, 由于网络流量的长相关性, 预测模型所需的距离相关训练样本的数量也要多于时间相关的训练样本. 所以基于距离相关的局域预测算法复杂度比全局算法低, 但比传统局域预测算法高.

表1 本文所提算法复杂度对比

No.	算法	计算量	平均耗时 (s)
1	局域LSSVM	$O\left(k_3'^3\right)$	0.037
2	分类LSSVM	$O\left(k_3''^3\right)+1\sim O\left(k_3''^3\right)+\left(3m+k_3''\right)n$	0.923
3	相关局域LSSVM	$O\left(k_3^3\right)+1\sim O\left(k_3^3\right)+(3m+k_2)n$	0.834
4	传统LSSVM	$O\left(n^3\right)$	

表2 小尺度网络流量预测算法性能对比

No.	算法	NMSE	文献
1	柔性神经树 + 粒子群	1.1215×10^{-2}	[14]
2	前馈神经网络模型 + 粒子群	7.32×10^{-2}	[14]
3	局域相关向量机 + 粒子群	8.5×10^{-3}	[16]
4	局域线性预测	1.43×10^{-2}	[16]
5	局域支持向量机	1.0135×10^{-2}	[15]
6	局域最小二乘支持向量机 + 模式搜索	1.18×10^{-2}	
7	改进的分类最小二乘支持向量机 + 模式搜索	5.6×10^{-3}	
8	相关局域最小二乘支持向量机 + 模式搜索	3.5×10^{-3}	

最后本文所提算法的预测精度与传统局域算法及新型改进算法的性能对比如表2所示.

表2中的局域最小二乘支持向量机和改进的分类最小二乘支持向量机两种算法的参数通过本文提出的参数优化方法寻优得到. 局域最小二乘支持向量机优化参数为$m'=7$, $C'=100000$, $\delta'^2=17044$, $\tau'=1$, $k_3'=47$, 可以从表2中看出LSSVM和SVM的预测性能接近, 也间接验证了LSSVM是一种性能优良的SVM改进算法, 在降低复杂度的同时仍然能够保持性能损失较小. 改进的分类最小二乘支持向量机的优化参数为$m''=7$, $C''=3839$, $\delta''^2=35$, $\tau''=1$, $k_3''=100$, 其性能要好于传统的局域预测算法, 说明在小尺度网络流量序列中, 利用自相似特性要好于利用时间相关性进行预测, 也证实了小尺度网络流量具有长相关和自相似的特性, 而其时变性较强, 传统局域预测难以适应其快速变化的特性. 从表2中可以看出, 本文所提的算法在预测精度上较好, 既利用了序列的自相似特性也利用了时间上的短相关, 预测性能相对现有最好算法增幅超过100%.

7 结 论

本文通过混沌理论分析了网络流量数据的主要特征, 在此基础上提出了一种基于相关分析的局域LSSVM 小尺度网络流量预测算法, 通过相关分

析同时选择时间相关训练子集和距离相关训练子集, 利用了网络流量时变性, 长相关和自相似特性. 相比于传统的局域预测模型, 本文所提算法不仅利用时间上邻近的训练样本, 同时在全部历史数据中搜寻欧氏距离最近的样本数据. 同时利用留一交叉验证法和模式搜索算法对训练子集中的距离相关样本和时间相关样本的数量比例进行了优化, 选择了最优子集. 通过对去噪后的小尺度网络流量趋势序列的预测实验证明, 本文所提算法不仅在同等条件下优于传统的局域预测算法或者分类预测算法, 而且在预测精度上好于现有的各种改进后的小尺度网络流量预测算法, 预测精度上的性能增益超过100%. 实验表明了小尺度网络流量具有较强的长相关和自相似特性, 同时其时变性也较强. 此外本文给出了一种在先验信息仅有训练数据的条件下, 通过留一交叉验证法和模式搜索算法自动选择预测模型优化参数组合以及最优训练子集的方法, 简单高效, 具有较强的实用性.

参考文献

[1] Man C T, Wong S C, Jian M X, Zhan R G, Peng Z 2009 *IEEE Trans. on Int. Trans. Sys.* **10** 60

[2] Marco L, Matteo B, Paolo F 2013 *IEEE Trans. on Int. Trans. Sys.* **2** 871

[3] Ana M, Rivalino M, Autran M, Paulo R M M, Lucio B A 2011 *12th International Conference on Parallel and Distributed Computing, Applications and Technologies* Gwangju, Korea, October 20–22, **2011** 109

[4] Jun J, Symeon P 2006 *Computer Communications* **29** 1627

[5] Li R, Chen J, Liu Y, Wang Z 2010 *The Journal of China Universities of Posts and Telecommunications* **17** 88

[6] Manoel C, Young S J, Myong K J, Lee D H 2009 *Expert Systems with Applications* **36** 6164

[7] Eleni I V, Matthew G K, John C G 2005 *Transportation Research Part C* **13** 211

[8] Chang H, Lee Y, Yoon B, Baek S 2011 *IET Intell. Transo. Syst.* **6** 292

[9] Tigran T T, Biswajit B, Margaret O M 2012 *IEEE Trans. on Int. Trans. Sys.* **13** 519

[10] Bao R C, Hsiu F T 2009 *Expert Systems with Applications* **36** 6960

[11] Sun H L, Jin Y H, Cui Y D, Cheng S D 2009 *Chin. Phys. B* **18** 4760

[12] Liu X W, Fang X M, Qin Z H, Ye C, Miao X 2011 *J. Netw. Syst. Manage* **19** 427

[13] Bao R C, Hsiu F T 2009 *Applied Soft Computing* **9** 1177

[14] Chen Y H, Yang B, Meng Q F 2012 *Applied Soft Computing* **12** 274

[15] Meng Q F, Chen Y H, Peng Y H 2009 *Chin. Phys. B* **18** 2194

[16] Meng Q F, Chen Y H, Feng Z Q, Wang F L, Chen S S 2013 *Acta Phys. Sin.* **62** 150509 (in Chinese) [孟庆芳, 陈月辉, 冯志全, 王枫林, 陈珊珊 2013 物理学报 **62** 150509]

[17] Vapnik V N 1999 *The Nature of Statistical Learning Theory* (2nd Ed.) (New York, Springer)

[18] Sapankevych N I, Sankar R 2009 *IEEE Comput. Intell. Mag.* **4** 24

[19] Wang X D, Ye M Y 2004 *Chin. Phys.* **13** 454

[20] Sun J C, Zhou Y T, Luo J G 2006 *Chin. Phys.* **15** 1208

[21] Liu H, Liu D, Deng L F 2006 *Chin. Phys.* **15** 1196

[22] 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]

[23] Farmer J D, Sidorowich J J 1987 *Phys. Rev. Lett.* **59** 845

[24] Jawad N, Keem S Y, Farrukh N, Sieh K T, Syed K A 2011 *Applied Soft Computing* **11** 4774

[25] Cai C Z, Fei J F, Wen Y F, Zhu X J, Xiao T T 2009 *Acta Phys. Sin.* **58** S008 (in Chinese) [蔡从中, 裴军芳, 温玉锋, 朱星键, 肖婷婷 2009 物理学报 **58** S008]

[26] Huang T Y 2008 *Chinese Journal Of Computers* **31** 1200 (in Chinese) [黄天云 2008 计算机学报 **31** 1200]

[27] Ligang Z, Kin K L, Lean Y 2009 *Soft Comput.* **13** 149

A local least square support vector machine prediction algorithm of small scale network traffic based on correlation analysis*

Tang Zhou-Jin† Peng Tao Wang Wen-Bo

(*School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*)

(Received 15 January 2014; revised manuscript received 11 April 2014)

Abstract

Real-time monitoring and forecasting technology for network traffic has played an important role in network management. Effective network traffic prediction could analyze and solve problems before overload occurs, which significantly improves network availability. In this paper, after the vulnerability of traditional nonlinear prediction method in forecasting modeling is analyzed, the relevant local (RL) forecast which is based on correlation analysis and the parameter optimization method based on pattern search (PS) is introduced. Using the correlation analysis, the optimal training subset is chosen from time-and distance-correlated training samples. On this basis, the prediction model is established by LSSVM. Finally network traffic dataset collected from wired campus networks is studied for our experiments. And the results show that the relevant local LSSVM prediction method whose training set and parameters have been automatically optimized can effectively predict the small scale traffic measurement data, and RL-LSSVM traffic forecasting algorithm exhibits significantly good prediction accuracy for the data set compared with previous algorithm.

Keywords: network traffic prediction, chaos time series forecasting, least squares support vector machine, local prediction

PACS: 05.45.Tp, 05.45.Gg **DOI:** 10.7498/aps.63.130504

* Project supported by the Chinese Defence Advance Research Program of Science and Technology, China (Grant No. 208010201).

† Corresponding author. E-mail: tangzhoujin@gmail.com