时间序列数据流的自适应预测

王永利[1,2,3] 周景华[4] 徐宏炳[1] 董逸生[1] 刘学军[1]

摘 要 提出一种自适应预测方法 AFStreams, 综合了复杂人工智能预测方法和时间序列预测方法的优点, 可以根据数据流值变化的快慢程度自适应地确定预测步长, 在计算资源受限的前提下, 形成最佳预测点轨迹. 仿真实验证明, AFStreams 能够良好地适应数据的变化, 在计算复杂度和预测精度之间平衡, 显著地提高了平均预测精度.

关键词 时间序列, 数据流, 预测, 插值小波, Kalman 滤波
中图分类号 TP274+.2; TP311.13

An Adaptive Forecasting Method for Time-Series Data Streams

WANG Yong-Li[1,2,3] ZHOU Jing-Hua[4] XU Hong-Bing[1] DONG Yi-Sheng[1] LIU Xue-Jun[1]

Abstract An adaptive forecasting method that combines the merits of the precision of artificial intelligence forecasting method and the rapidness of times-series forecasting method, called AFStreams, is proposed. It can estimate the forecasting-step self adaptively from the change ratio of stream-values and can generate proved optimal track of forecasting points with the minimum computation cost from limited resources. Experiments proved that AFStreams can adapt to the changes of data well and provide tradeoff between computing complexity and forecasting precision.

Key words Time-series, data streams, forecasting, interpolating wavelet, Kalman filtering

1 引言

数据流技术[1]近年来得到了越来越多的关注, 主要的研究方向包括资源受限计算[2], 模式发现[3,4], 数据流挖掘[5]等. 目前数据流研究领域中已有的趋势分析理论和方法[2~4]一般关注于相似性或模式差异的预测, 有关数据流值本身预测的文献较少.

由于数据流可以视为时间序列[1], 来自于控制理论的一些思想越来越多地被引入到近似计算[2]和数据流挖掘[5] 研究领域. 多数应用产生的离散数据流 $\boldsymbol{s}[t]$ 的数学模型可以简化为[6] : $\boldsymbol{s}[t] = \boldsymbol{a}[t] + \boldsymbol{w}[t]$, 其中 $\boldsymbol{a}[t]$ 表示稳定性成分, $\boldsymbol{w}[t]$ 表示随机噪声. $\boldsymbol{a}[t]$ 的变化相对有规律, 只受少量相关因素影响, 因而可以精确预测, 但预测周期一般较长. 时间序列领域已经取得了许多有关预测的研究成果, 基本上可以提供 $\boldsymbol{a}[t]$ 的精确预测. 然而由于随机成分 $\boldsymbol{w}[t]$ 通常容易受到各种随机因素的影响, 怎样精确地预测 $\boldsymbol{w}[t]$ 就成为数据流预测研究中的关键问题.

已经提出的预测方法[2,5,7,8]大都没有和精确的稳定性成分预测算法相结合, 预测的精度有待进一步提高, 而且它们侧重点只是研究如何提高预测精度, 忽略了在计算资源受限的条件下, 预测计算代价与预测精度的平衡问题.

本文提出一种自适应预测精度与计算复杂度的数据流值预测方法–AFStreams (Adaptive forecasting method for stream-values), 其基本思想是: 利用相对精确的稳定成分预测方法得出的长间隔预测信息, 在实际测量值与最近一个长间隔预测值之间, 根据数据流值的变化情况, 以短的间隔插值, 同时不断修正实际测量值与长间隔预测值和实际测量值与短间隔预测插值的误差, 自适应地平衡预测精度和计算复杂度.

2 模型概述

本文中, 术语"流值"表示单数据流中数据项的单个属性值; 术语"步长"表示流的预测值到当前观测值之间的时间间隔; 术语"尺度"表示小波插值的分辨率. 为实现对数据流上两种成分的连续预测, 引入"双滑动窗口"结构: 长度为 I 的窗口称为"长窗口", 用作稳定成分预测的基本更新单位; 小窗口称为"短窗口", 用作随机成分预测的基本更新单位. 流值 $\boldsymbol{s}[t]$ 的预测开始于 t 时刻, 每隔 Δt 秒采样一次, 假设采样间隔 Δt 恒定. 长窗口的当前时刻用 t_1 表示, 以 I 为单位更新; 短窗口的当前时刻用 t_2 表示,

收稿日期 2005-5-5 收修改稿日期 2006-6-29
Received May 5, 2005; in revised form June 29, 2006
江苏省研究生创新计划项目 (xm04-36) 资助
Supported by Graduate Creative Program Foundation of Jiangsu (xm04-36)
1. 东南大学计算机科学与工程学院 南京 210096 2. 佳木斯大学公共计算机教研部 佳木斯 154007 3. 南京理工大学计算机科学与技术学院 南京 210094 4.上海伽兴电子科技有限公司 上海 200233
1. School of Computer Science and Engineering, Southeast University, Nanjing 210096 2. Department of Common Computer Teaching, Jiamusi University, Jiamusi 154007 3. School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094 4. Jia Xing Electronics Technology Co., Ltd. Shanghai 200233
DOI: 10.1360/aas-007-0197

以 $k\Delta t$ 为单位更新. AFStreams 预测模型的结构如图 1 所示, 由 5 个功能模块组成:

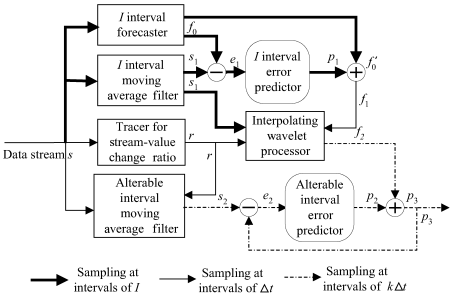

图 1 自适应精度数据流值预测模型结构图

Fig. 1 Adaptive precision forecasting data streams values model in variable forecasting-step

1) 移动平均滤波器. 用于滤除原始数据流中的噪声, 分为长间隔与短间隔两种, 每次窗口平移时对窗口中的数据取平均值作为当前最新的测量值.

2) 长间隔预测器. 用于精确预测数据流中的稳定性成分, 两个相邻稳定性成分预测值的时间间隔为 $I\,(I > k\Delta t)$, 使用较为耗时但精度较高的预测方法 $\mathrm{I}_{forecast}$, 如神经网络、支持向量机等, 条件是 $O(\mathrm{I}_{forecast}) < I$.

3) 流值变化跟踪器. 用于量化流值变化的强度, 定义数据流序列 s 在 t 处的流值变化率为 $\delta(t) = (\boldsymbol{s}(t) - \boldsymbol{s}(t-1))/\Delta t$. 为了排除偶然因素, 设 $\bar{\delta}$ 为变化率的加权平均值, 流值变化跟踪器以 Δt 为采样间隔捕获 $\bar{\delta}$, 根据 $\bar{\delta}$ 映像成不同的预测步长级别.

4) 小波插值器. 用于生成具有更加精细的预测步长的预测流值, 最大插值尺度 $r_{\max} = \ulcorner \log_2(I - t + 1/\Delta t) \urcorner$. 为了映射流值变化与插值尺度 (直接决定预测步长) 之间的关系, 定义尺度导引算子 $\Gamma_t(\bar{\delta}(t)) : A_t(\Omega) \rightarrow V_j(\Omega)$, 其中 $A_t(\Omega)$ 表示流值变化率空间, $V_j(\Omega)$ 表示尺度空间, 本文中设 $\Gamma_t(\bar{\delta}(t)) = \bar{\delta}(t) \cdot r_{\max} \,/((\max(\boldsymbol{s}[t]) - \min(\boldsymbol{s}[t]))/\Delta t)$.

5) 误差预测子. 用于对最近一次预测值及时修正, 根据当前状态和预测值与实际测量值之间的误差, 估计未来的误差. 误差预测子由长间隔误差预测子和短间隔误差预测子组成. 前者负责估计慢速移动的低频流值成分, 后者主要估计快速移动的高频流值成分.

模型的预测过程是: 数据流 s 由长间隔预测器提供未来两个长间隔预测值 f_0, f_0', 待 t 超过 t_1 之后, 长间隔误差预测子以 $e_1 (e_1 = s_1 - f_0)$ 为输入, 形成误差估计 p_1, 进而形成改良的长间隔预测值 $f_1 (f_1 = p_1 + f_0')$, 然后小波插值器在流值变化跟踪器的导引下 (对于定长短间隔情况, 不需考虑流值变化), 以短间隔 r 在 s_1 和 f_1 二进插值, 产生预测插值 f_2, 短间隔误差预测子以 $e_2 (e_2 = s_2 - f_2', f_2'$ 应为前一次预测得出的最终预测值 p_3) 为输入, 形成短间隔误差估计 p_2, 与 f_2 生成最后的预测流值 $p_3 (p_3 = p_2 + f_2)$. 模型中包含的两个预测子使用 Kalman 滤波构造.

3 理论依据

3.1 二进插值小波

定义 1 (带有尺度导引的插值小波). 对于 $f \neg \in B_j$, 在二进空间 B_j 上定义投影子, 它对二进采样 $f(2^j k)$ 进行插值, 插值公式[9]为:

$$P_{B_j} f(t) = \sum_{k=-\infty}^{+\infty} f(2_j k) \phi_j (t - 2^j k) \qquad (1)$$

$P_{B_j} f(t)$ 为 $p-1$ 阶插值多项式, ϕ 为插值函数, 插值函数的伸缩产生一个不同采样间隔下的新插值, 最大插值尺度 j 依赖于尺度导引算子 $\Gamma_t(\bar{\delta}(t))$, $\phi_j (t - 2^j k)_{(k \in Z)}$ 是其生成空间的一组 Riesz 基. 称 $\psi_{j,k} = \phi_{j-1,2k+1}$ 为带有尺度导引的插值小波.

根据二进插值小波的构造特性, 在一个给定的区间上产生相同间隔的插值点, 二进插值比等间隔插值代价小. 这是 AFStreams 模型在可变预测步长情况下采用二进插值的主要依据. 预测步长与插值尺度 (二进的次数) 有直接关系, 预测步长 (插值间隔) 越小, 插值代价越大.

3.2 Kalman 滤波误差预测子

Kalman 滤波 (KF) 能够给出状态的一步最小均方误差估计 (Minimum mean-square error, MMSE)[10]. 为了实现可变预测步长的预测误差的修正, 我们定义一种允许观测不相邻的采样值的特殊 Kalman 滤波, 称为 SKF (Special Kalman filtering).

定义 2. 不失一般性, 对于任意线性离散时间系统, 适用于一般流应用的矢量 SKF 模型定义如下: 状态模型: $\boldsymbol{x}_{n+1} = \Phi_n \boldsymbol{x}_n + D_n \boldsymbol{u}_n$, 测量模型: $\boldsymbol{z}_n = H_n \boldsymbol{x}_n + \boldsymbol{w}_n$.

其中, $\boldsymbol{x}_{n+1}$ 为 p 维状态向量, Φ_n, D_n 是已知的 $p \times p$ 和 $p \times r$ 的状态转移矩阵, $\boldsymbol{u}_n$ 是服从 $\boldsymbol{u}_n \sim N(0, Q_n)$ 的 p 维输入高斯白噪声, $\boldsymbol{x}_{-1} \sim N(\boldsymbol{u}_s, C_s)$, $\boldsymbol{x}_{-1}$ 与 $\boldsymbol{u}_n$ 相互独立, $\boldsymbol{z}_n$ 为 m 维的测量向量, H_n 为 $m \times p$ 维测量矩阵, $\boldsymbol{w}_n$ 是服从 $\boldsymbol{w}_n \sim N(0, R_n)$ 的 m 维测量高斯白噪声.

状态 $\boldsymbol{x}_n$ 的估计 $\hat{\boldsymbol{x}}_n = \hat{\boldsymbol{x}}_{n-1} + \boldsymbol{k}_n (\boldsymbol{z}_n - H \hat{\boldsymbol{x}}_{n-1})$, $\boldsymbol{z}_n - H \hat{\boldsymbol{x}}_{n-1}$ 表示测量值和其估计的差值, $\boldsymbol{k}_n$ 表示每次调整的权值, 称为 Kalman 增益; $\boldsymbol{z}_n$ 的预测值为 $H \hat{\boldsymbol{x}}_n + \boldsymbol{v}_n$.

当 $\Phi_n = a$ (a为常数), $H_n = I$, $\boldsymbol{w}_k$ 服从 $\boldsymbol{w}_k \sim N(0, \sigma_w^2)$, 状态量和观测量都为标量的时候, 表示标

量 SKF 模型; 当 Φ_n, H_n 中的变量全部为常数时, 表示常量矢量 SKF; 当 Φ_n, H_n 中的变量为时间 n 的函数时 (即时变的), 表示状态或观测方程都是非线性序贯状态的估计, 表示扩展 SKF 模型.

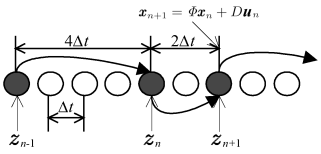

图 2　可变测量值间隔的特殊 Kalman 滤波 (SKF) 示意

Fig. 2　Special Kalman filtering with variable intervals for measured values

可变测量值间隔的 SKF 如图 2 所示, 阴影圆点表示相邻的三个 SKF 测量值, 每个状态空间蕴含了多个采样点的信息. 下面讨论状态估计的均方误差与预测步长的关系.

引理 1. SKF 模型中, 设 t_i 和 t_j 表示从时刻 t 开始以不同的预测步长 i 和 j 预测的两个未来时刻, M_i 与 M_j 表示两时刻的状态的最小均方误差估计, 如果 $i{\leq}j$, 那么 $M_i{\leq}M_j$.

证明. 如果一步 SKF 跨越了 n 个观测量和状态量, 我们使用 n 步标准 KF 递推式构造一步 SKF 递推式. 对于标准 KF, 设 $M_{n|n-1}$ 表示在前 $n-1$ 基础上第 n 个状态的一步 MMSE, 最佳状态估计的准则就是使贝叶斯均方误差估计 $M_{n|n-1}=\mathrm{E}[(\boldsymbol{x}_n-\hat{\boldsymbol{x}}_{n|n-1})^2]$ 的误差最小. 为叙述方便, 考虑标量 KF 情况, 其 MMSE 的迭代式[10]为: $M_{n|n-1}=a^2M_{n-1|n-1}+\sigma_u^2$.

由于在标准标量 KF 中, Kalman 增益为 $K_n=M_{n|n-1}(\sigma_u^2+M_{n|n-1})^{-1}$, MMSE为 $M_{n|n}=(1-K_n)M_{n|n-1}$, 已知 $a{\leq}1$, $M_{-1|-1}=\sigma_s^2$, 有: $M_{n|n}=(1-K_n)M_{n|n-1}=(1-\frac{M_{n|n-1}}{\sigma_n^2+M_{n|n-1}})M_{n|n-1}=\frac{\sigma_n^2M_{n|n-1}}{\sigma_n^2+M_{n|n-1}}$, 于是 $M_{i|i-1}=a^2M_{i-1|i-1}+\sigma_u^2=a^2(\frac{\sigma_n^2M_{i-1|i-2}}{\sigma_n^2+M_{i-1|i-2}})+\sigma_u^2=\frac{a^2\sigma_n^2}{\frac{\sigma_n^2}{M_{i-1|i-2}}+1}+\sigma_u^2=\frac{a^2\sigma_n^2}{\frac{\sigma_n^2}{a^2M_{i-2|i-2}+\sigma_u^2}+1}+\sigma_u^2=\cdots$, 由于 $\frac{\sigma_n^2}{M_{i-1|i-2}}+1$ 是 i 的减函数, 那么 $M_{i|i-1}$ 就是 i 的增函数, 所以$M_{i|i-1}{\leq}M_{j|j-1}$. □

对于矢量 SKF 模型, $M_{n|n-1}=\Phi M_{n-1|n-1}\Phi^{\mathrm{T}}+DQD^{\mathrm{T}}$, 可以有类似的推导. 引理 1 说明预测的误差具有累加效应. SKF 中相邻的测量点之间包含的采样信息越多, MMSE 越大, 预测精度越低.

3.3　最佳预测点轨迹的确定

定义 3. 预测精度的 ε 近似. 对于所有可能的预测步空间 A, 设 M_i 与 M_j 分别为预测步 i 与 j 的 MMSE, 如果 $\sup_{i,j\in A}|M_i-M_j|\leq\varepsilon$ 成立, 称预测步 i 与 j 在 A 空间中的预测精度是 ε 近似的.

定义 4. 最佳预测点轨迹. 令 c_i 表示构造第 i 个预测步的代价, 在与用户指定的预测步约束满足预测精度 ε 近似的前提下, 选择合适的预测点可以令生成每个预测值的平均代价 $\bar{c}$ 最小, 且能够适应流值的变化情况. 称由多个这样预测点组成的序列为最佳预测点轨迹.

定理 1. 在保证指定预测精度的前提下, 平均计算代价最低的最佳预测点轨迹存在并且可解.

证明. 选择一组变化时刻集合 $n_0,n_1,...,n_{N_s-2}$ (定义 n_{-1} 和 $n_{N_s-1}=N$), 将两个 I 预测间隔之间划分为 N_s 段, 希望确定一个可以满足约束条件且平均计算代价最低的预测时刻序列, 这是一个动态规划 (DP) 问题.

目标函数: $\min C=\sum_{i=1}^{k}c_i$ ($c_i=c_i^s+c_i^p$, c_i^s 表示计算预测步开销, c_i^p 表示预测开销).

约束条件: 设 l_u 为用户指定的最小预测步长, i 为 DP 选定预测步长, 要求 l_u 与 i 预测精度是 ε 近似的, 并且插值尺度 $r_i\in[l_{\max}-\ulcorner\log_2 l_u\urcorner,l_{\max}]$.

最佳分段是使目标函数 $\sum_{i=0}^{N_s-1}c_i||M_i-M_{l_u}|<\varepsilon$ 最小的那些值. 为了应用动态规划法, 定义 $\Delta_i[n_{i-1},n_i-1]=\min c_i$, 希望使 $\sum_{i=0}^{N_s-1}\Delta_i[n_{i-1},n_i-1]$ 最小.

定义 $I_k[L]=\min_{n_0,n_1,...,n_{k-1}}\sum_{i=0}^{k}\Delta_i[n_{i-1},n_i-1]$, 其中 $n_{-1}=0$, $0<n_0<n_1<...<n_{k-1}<L+1$, $n_k=L+1$, 建立一个最小值的递归计算方法.

$$I_k[L]=\min_{n_{k-1}}\min_{n_0,n_1,...,n_{k-2}}\sum_{i=0}^{k-1}\Delta_i[n_{i-1},n_i-1]$$

$$=\min_{n_{k-1}}\min_{n_0,n_1,...,n_{k-2}}\sum_{i=0}^{k-1}\Delta_i[n_{i-1},n_i-1]+\Delta_k[n_{k-1},n_k-1]$$

$$=\min_{n_{k-1}}[(\min_{n_0,n_1,...,n_{k-2}}\sum_{i=0}^{k-1}\Delta_i[n_{i-1},n_i-1])+\Delta_k[n_{k-1},n_k-1]]$$

最后有 $I_k[L]=\min_{n_{k-1}}(I_{k-1}[n_{k-1}-1]+\Delta_k[n_{k-1},L])$,即对于数据元组 [0,$L$], k+1(k个变化时刻) 段的最小值, 是在 $n=n_{k-1}-1$ 结束的前 k 段最小值与从 $n=n_{k-1}$ 到 $n=L$ 的最后一段值之和.

按照插值尺度映射得到 l, 在其基础上左延, ..., l-2, l-1, l, 找到最小一个满足 $|M_{l'}-M_{ls}|<\varepsilon$ 的尺度 l' 作为最佳插值尺度 r_i, 则能够令第 i 段代价 $\Delta_i[n_{i-1},n_i-1]$ 最小, 逆向应用上述方法, 最佳预测点轨迹存在, 并且可以通过动态规划法在多项式时间内找到局部最优方案. □

AFStreams 预测模型复杂度分析: I 间隔预测器可以应用最先进的技术获得精度尽可能高的 I 间隔预测, 其运行时间保证在 I 间隔之内完成即可, 不列入复杂度估算中; 估计 I 间隔误差和估

计小间隔多尺度误差的时间代价与应用中选取的状态变量规模有关. 对于一般矢量 SKF, 设状态矢量中有 p 个变量, 测量矢量中有 m 个变量 (通常$m \leq p$), 则一步误差的预测代价为 $O(p \times m)$; 流值变化率跟踪器采用增量方式计算平均变化率 $\bar{\delta}(t)$ (即$\bar{\delta}(t+1) = \bar{\delta}(t) + \delta(t+1) - \delta(t-n+1)$), 一步计算最佳插值尺度 r 的时间代价为$O(1)$; 判断精度是否满足约束时, MMSE 矩阵的计算与状态矢量有关, 时间代价为 $O(p^2)$; 多尺度小波插值根据文献 [9] 时间代价为 $O((\log_2(I/\Delta t))^2)$; 除了上述几步, 其它步骤均可以在常数时间内完成, 因此算法的一次预测总体时间代价为 $O((\log_2(I/\Delta t))^2)$ 或者 $O(p \times m)$. 对于大多数应用, 模型的时间复杂度为 $O((\log_2(I/\Delta t))^2)$. 对于经过了归一化和坏数据修补等预处理的数据流值, $I/\Delta t$ 的取值范围相对稳定, 与数据项的数目无关, 因此 AFStreams 预测模型适合处理大规模数据集.

4 仿真实验及分析

实验数据来自于加州大学时间序列研究中心[11]提供的某地区 1995~1998 年电力负荷真实数据集, 其中包含 29932 条负荷记录, 1309 条天气记录, 共 5.10M, 负荷采样间隔为 5min. 稳定性 I 间隔预报值 (I=1hr) 使用天气敏感型 BP 神经网络预测. 预处理时, 规格化负荷数据流和温度数据流序列 x_1, x_2, x_3,$\cdots$, x_N的方法是: $x'_i = (x_i - m_x)/s_x$, 其中 m_x 为算术均值, s_x 为标准方差.

我们应用 Welch's 方法[12]分析两种误差 (长间隔与短间隔) 的能量频谱密度 (Power spectral density, PSD) 特征, 通过匹配误差 PSD 幅度频率, 反应采样集频谱内容的方法构造两个误差 KF 模型的参数. 经过对 1997 年 6~8 三个月数据反复地频域识别和频谱分析, 得到一个三阶的长间隔误差标准 KF 模型, 两个二阶的短间隔误差 SKF 模型. 限于篇幅,模型的细节省略.

实验 1 测试 AFStreams 引入稳定性成分预测的效果. 我们同时实现了单纯基于 KF 的预测方法 (PureKF)[2] 和线性回归预测方法 (Linear regression forecasting, LRF)[5], 与 AFStreams 在固定预测步长的情况下, 对比三种算法在 5min, 10min 等几个步长级别的预测精度. 定义 i 预测点的相对预测误差 $\Delta p_i = (v_i^f - v_i^a)/v_i^a \times 100\%$, 其中 v_i^f 表示预测流值, v_i^a 表示实际测量流值. 我们对不同的 20 天负荷数据进行了 20 次测试, 取每一级别的平均相对误差, 实验结果如图 3 所示. AFStreams 的预测误差在 5, 10, 15 级别时很小, 其平均预测精度明显高于 PureKF 和 LRF. PureKF 的预测误差随着步长的加大而加大, 这与 KF 的性质有关. LRF 在步长加大的时候仍然有良好表现. 总的来看, AFStreams 比另外两种预测方法的相对误差平均低 1% 甚至 3%, 精度较高, 算法运行稳定. 引入稳定性成分显著地提高了平均预测精度.

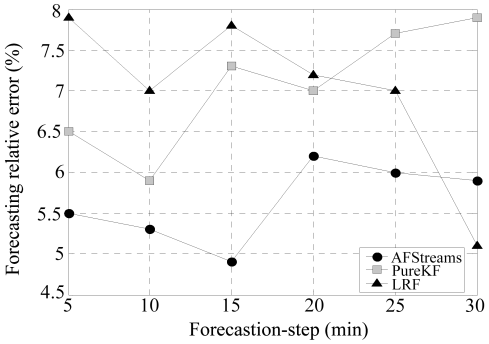

图 3 5min 级相对预测误差比较

Fig. 3 Comparison of relative error values

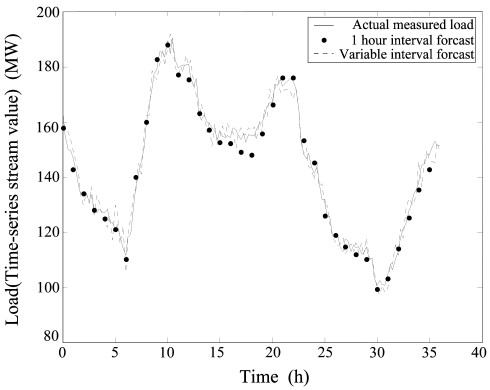

图 4 AFStreams 的 36hr 预测值与实际值曲线

Fig. 4 36 hours curve

实验 2 测试 AFStreams 在自适应预测步长情况下的平均预测精度. 本文以 1997 年 7 月 5~6 日的 36 小时预测结果为例分析实验结果 (由于在 5 日发生了一场大暴雨, 降雨引起的抗洪, 排渍负荷是稳定性预测方法难以预测的, 是测试随机成分预测精度的合适场景). 用户指定最大步长 l_u=20min (即自适应预测步不得超过 20min), 精度约束 ε=0.1. 图 4 显示了预测流值与实际流值的性能曲线, 共产生 386 个小间隔预测点 (1 小时的平均预测点为 10.72 个), 产生每个小间隔预测的平均时间为 47.4ms.

为比较可变和固定预测步长的性能, 定义 $\bar{P}$ 为平均预测精度 ($\bar{P}$=$[1 - \sqrt{\frac{1}{n}\sum_{i=1}^{n} \Delta p_i^2}] \times 100\%$), 自适应预测步长时 $\bar{P}$=94.566%, 而 15min 固定步长时 $\bar{P}$=93.21%, 5min 固定步长时 $\bar{P}$=93.6%. 由于小间隔预测子能够感知数据变化特征, 负荷变化剧烈的时段预测精度并没有较大变化, 引入多尺度插值点的精度改良作用得到了验证.

36 小时的预报结果证明, AFStreams 无论在拐点还是在负荷快速波动时段都有较好的适应能力.

这是因为 AFStreams 以误差而不是原始的数据作为 KF 的状态向量和测量向量, 更好地反映了时变系统的变化规律.

5 结论

本文从典型应用出发, 深入地分析了数据流值预测的问题, 所提出的 AFStreams 方法综合了非线性预测方法精确和线性预测方法快速的优点, 相比于传统的预测方法主要有两个改进之处：确保了更好的精度并且提供了更加丰富的预测步长. 简单地说, 就是在应该精确预测的地方加细预测步长, 在可以节省计算代价的地方拉长预测步长. 下一步工作包括研究新的自适应预测尺度的方法, 进一步提高预测精度, 研究基于硬件的可实际应用的数据流处理系统.

References

1 Babcock B, Babu S, Datar M, Motwani R, Widom J. Models and issues in data stream systems. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Madison, Wisconsin, USA: ACM Press, 2002. 1∼16

2 Jain A, Chang E Y, Wang Yuan-Fang. Adaptive stream resource management using Kalman filters. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. Paris, France, USA: ACM Press, 2004. 11∼22

3 Papadimitriou S, Sun Ji-Meng, Faloutsos C. Streaming pattern discovery in multiple time-series. In: Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, Norway: VLDB Endowment, 2005. 697∼708

4 Sun Ji-Meng, Papadimitriou S, Faloutsos C. Distributed pattern discovery in multiple streams. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD). Singapore, Berlin: Springer-Verlag, LNCS, 2006. **73918**: 13∼718

5 Faloutsos C. Stream and sensor data mining. In: Proceedings of the 9th International Conference on Extending Database Technology. Heraklion, Greece, Berlin: Springer-Verlag, LNCS, 2004. 25∼27

6 Trudnowski J D, McReynolds W L, Johnson M J. Real-time very short-term load prediction for power-system automatic generation control. *IEEE Transactions on Control Systems Technology*, 2001, **9**(2): 254∼260

7 Liu K, Subbarayan S, Shoults R R, Manry M T, Kwan C, Lewis F I, Naccarino J. Comparison of very short-term load forecasting techniques. *IEEE Transactions on Power Systems*, 1996, **11**(2): 877∼882

8 He Guo-Guang, Ma Shou-Feng, Li Yu. A study on forecasting for time series based on wavelet analysis. *Acta Automatica Sinica*, 2002, **28**(6): 1012∼1014 (in Chinese)
(贺国光, 马寿峰, 李宇. 基于小波分解与重构的时间序列预测法. 自动化学报, 2002, **28**(6): 1012∼1014)

9 Mallat S. *A Wavelet Tour of Signal Processing, Second Edition*. Boston: Academic Press, 1999. 221∼226

10 Brown R G, Hwang P Y C. *Introduction to Random Signals and Applied Kalman Filtering*, 2nd Edition. New York: John Wiley&Sons, 1992. 134∼168

11 Keogh E, Folias T. The UCR Time Series Data Mining Archive [Online], available: http://www.cs.ucr.edu/∼eamonn/TSDMA/index.html, Riverside CA: University of California-Computer Science & Engineering Department, 2002

12 Oppenheim A V, Schafer R W, Buck J R. *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989. 301∼345

王永利 博士, 讲师, 研究领域为数据流分析, 现代数据库技术, 模式识别. 本文通信作者. E-mail: wyl_seu@126.com
(**WANG Yong-Li** Ph. D. in Database Laboratory at Southeast University, lecturer. His research interests include data streams analyzing, model database technique, and pattern recognition. Corresponding author of this paper.)

周景华 工程师, 上海伽兴电子科技有限公司总经理, 研究领域为工业自动化控制, 数据处理. E-mail: jinghua.zhou@asiacontrol.com.cn
(**ZHOU Jing-Hua** Engineer in Department of Research, Jiaxing Electronic Technology Ltd. Co.. His research interests include industry automatic control and data processing.)

徐宏炳 教授, 研究领域为系统结构，现代数据库技术，数据流管理. E-mail: hbxu@seu.edu.cn
(**XU Hong-Bing** Professor in Southeast University. His research interests include system architecture, model database technique, and data streams management.)

董逸生 教授, 研究领域为现代数据库技术，信息系统建模，生物信息学. E-mail: ysdong@seu.edu.cn
(**DONG Yi-Sheng** Professor in Southeast University. His research interests include model database technique, designing and modeling information system, and bioinformatics.)

刘学军 博士, 讲师, 研究领域为数据流挖掘, 现代数据库技术.
E-mail: lxj-gd@sina.com.cn
(**LIU Xue-Jun** Ph. D. in Database Laboratory at Southeast University, lecturer. His research interests include data streams mining and model database technique.)