一种数据流趋势分析方法的研究与应用①

汪成亮 [1,2]　陆志坚 [1]　庞　栩 [1]

(1.重庆大学 计算机学院 重庆 400044; 2.重庆大学 电器工程学院 重庆 400044)

摘　要： 有效趋势的提取可为监控对象提供早期预警、状态评估和决策支持。传统的曲线趋势分析算法有滑动窗口(SW)算法、外推式在线数据分割(OSD) 算法，二者均采用常规最小二乘法进行曲线拟合。与常规最小二乘法相比，总体最小二乘法具有更高的直线拟合精度。此外，针对 SW 算法的滑动窗口最大长度没有限制，当检测点阈值比较大时，窗口的长度可能很长；而 OSD 算法限定了最小滑动窗口长度，使得在最小滑动窗口内的突变点无法检测。针对 SW 算法和 OSD 算法的缺陷，提出了一种新的数据流趋势分析方法，该方法采用总体最小二乘法对数据流进行分段拟合，提高了趋势分析精度；还提出了可变滑动窗口算法解决 SW 算法和 OSD 算法的固定窗口问题，以实现对数据流的合理分割。实验结果表明，有效性较为明显。

关键词： 总体最小二乘法 可变滑动窗口 趋势分析 动态数据挖掘

Research and Application of an Algorithm for Trend Analysis of Data Streams

WANG Cheng-Liang[1,2], LU Zhi-Jian[1], PANG Xu[1] (1.College of Computer Science , Chongqing University, Chongqing 400044, China; 2. College of Electrical Engineering , Chongqing University, Chongqing 400044, China)

Abstract: Efficient trend extraction methods can provide early warnings, severity assessments of monitored subjects and information for decision support. The traditional algorithms for trend analysis of curves include Sliding Window algorithm (SW) and Extrapolation for On-line Segmentation of Data algorithm (OSD), which use total least squares for curve fitting. Compared with conventional least squares, the total least squares has a higher accuracy of fitting a straight line. In addition, since there is no restriction on the maximum length of the sliding window for SW algorithm, the length of window can be very long when threshold for Detection of point becomes larger. As OSD algorithm restricts the minimum length of sliding window, mutations within minimum sliding window cannot be detected for defects of the SW algorithm and the OSD algorithm. This paper presents a new method for trend analysis of data streams. The method uses total least squares to improve the accuracy of trend analysis. It also presents variable sliding window algorithm to solve the fixed window problem with the SW algorithm and OSD algorithm to achieve a reasonable segmentation for data streams. The experimental results show that the method is effective.

Keywords: total least squares; changeable sliding window; trend analysis; dynamic data mining

数据流是连续的、无限的、快速的、随时间变化的数据元素组成的序列，是一种特殊的数据类型。它在一个近似无限长的时间范围内，快速产生大量数据[1]。动态数据流趋势分析的目的是提取趋势变化信息，为监控对象提供早期预警、状态评估和决策支持[2]。由于应用领域的不同，动态数据流自身曲线特征千差万别，因此

① 基金项目:重庆市自然科学基金(CSTC)(2007BB6118);中国博士后科学基金(20080430750)
收稿时间:2009-05-04

要求数据流趋势分析方法适应性强、分析精度高，能够应用于各个领域的动态数据流分析、预测。文章提出了一种新的数据流趋势分析方法，该算法采用总体最小二乘法对数据流进行分段拟合，提高了趋势分析精度；采用可变滑动窗口算法实现对数据流的合理分割。实验结果表明，有效性较为明显。

1 数据流问题描述

为了研究需要，我们定义：

不断到达的一维时序数据流为

$$Y=\{v_{t_1},\ldots,v_{t_i},\ldots,v_{t_c},\ldots\} \tag{1}$$

其中，t_c 为当前时刻。

数据流分割(本文采用均方差比较)将 Y 分割成一系列连续的非空数据段(即滑动窗口)：

$$\{Y_1,\ldots,Y_j,\ldots,Y_s,\ldots\} \tag{2}$$

其中第 j 数据段为：

$$Y_j=\left\{v_{t_{j1}},\ldots,v_{t_{j\lambda}},\ldots,v_{t_{jnj}}\right\} \tag{3}$$

对应的数据到达时间：

$$t_{j,\lambda}\in\{t_1,\ldots,t_i,\ldots,t_c,\ldots\}(j\in N,1\le j\le s;\lambda\in N,1\le\lambda\le n_j) \tag{4}$$

在(3)式和(4)式中，我们用 n_j 表示数据段 Y_j 的长度，即滑动窗口 Y_j 的长度，而且令 $t_{1,1}=t_1$。

当前数据段 Y_s：包含当前数据 v_{t_c} 的数据段。

设 Y_j 中的数据可用总体最小二乘法进行拟合，即

$$v(t)=(a_j+\delta_j(t))t+(b_j+\varepsilon_j(t))(t\in\{t_{j,1},\ldots,t_{j,n_j}\}) \tag{5}$$

其中，a_j 和 b_j 为模型参数，参数 a_j 称为数据段 Y_j 的趋势特征值，$\delta_j(t)$ 为 a_j 的误差扰动，$\varepsilon_j(t)$ 为独立同分布零均值白噪声。

为了算法描述的需要，我们设数据段 Y_{j+1} 的第 1 个数据元素 $v_{t_{j+1,1}}$ 为 Y_j 的分割点。另外，令 β 为模型参数向量，

$$\beta=\left[a_j,b_j\right]^T \tag{6}$$

滑动窗口(SW) 算法是一种连续建模分析数据流分割算法，即在当前已建立回归模型的数据段基础上，用新到达的每一数据扩充当前数据段，并重新建立新的回归模型。若该模型的拟合均方差大于预先给定的分割点阈值，则认为新到达的数据为当前数据段的分割点，该数据段的趋势特征值由已有回归模型的参数给出，并将后续到达的数据归入新的当前数据段，启动新的趋势分析过程；若上述分割点检测判据不成立，则继续分析下一个到达的数据。Sylvie 等提出一次建模、连续分析的外推式在线数据分割(OSD) 算法。该算法待到当前数据序列达到一定长度时，才对其建立回归模型；此后对于新到达的数据，只将其代入已建立的模型，分析外推累积误差。 若大于事先给定的阈值，则认为新到达的数据为当前数据段的分割点，并从已建立的模型中获得该数据段的趋势特征值；若上述分割点检测判据不成立，则继续分析下一个到达的数据[3]。

滑动窗口(SW)算法、外推式在线数据分割(OSD) 算法，二者均采用常规最小二乘法进行曲线拟合。与常规最小二乘法相比，总体最小二乘法具有更高的直线拟合精度[4]。此外，针对 SW 算法的滑动窗口最大长度没有限制，当检测点阀值比较大时，窗口的长度可能很长使得趋势分析的误差变大；而 OSD 算法限定了最小滑动窗口长度，使得在最小滑动窗口内的突变点无法检测。

2 数据流趋势分析方法

针对现有趋势分析算法的缺陷，本文提出一种新的数据流趋势分析方法，用以克服 SW 算法和 OSD 的弊端。

2.1 总体最小二乘法回归建模

设当前数据序列 $Y_{s,n}=\{v_{t_{s,1}},v_{t_{s,2}},\ldots,v_{t_{s,n}}\}$ 构造的线性回归模型为：

$$v(t)=(a_{s,n}+\delta)t+b_{s,n}+\varepsilon \tag{7}$$

若该模型函数通过点 $(t_{s,i},v_{t_{s,i}})$，则(7)式可转化为：

$$v-v_{t_{s,i}}=a_{s,n}(t-t_{s,i}) \tag{8}$$

设 $(\bar{t},\bar{v})$ 为当前数据段 $Y_{s,n}$ 的几何中心点。其中 $\bar{t},\bar{y}$ 分别是：$\bar{t}=\frac{1}{n}\sum_{i=1}^{n}t_{s,i}$，$\bar{v}=\frac{1}{n}\sum_{i=1}^{n}v_{t_{s,i}}$

此时回归模型方程为：

$$v-\bar{v}=a_{s,n}(t-\bar{t}) \tag{9}$$

令中心化的数据向量为：

$$T=\left[t_{s,1}-\bar{t},t_{s,2}-\bar{t},\ldots,t_{s,n}-\bar{t}\right]^T,$$

$$V=\left[v_{t_{s,1}}-\bar{v},v_{t_{s,2}}-\bar{v},\ldots,v_{t_{s,n}}-\bar{v}\right]^T$$

常规最小二乘法极小化的范数 $\xi_{LS}=\|\varepsilon\|^2$，这等价于在条件 $T\cdot a_{s,n}=V-\varepsilon$ 的约束下，$\|\varepsilon\|^2$ 达到最小化。也就是使校正项 ε 尽可能小，同时通过强令 $T\cdot a_{s,n}=V-\varepsilon$ 补偿存在于向量 Y 中的噪声。T 和 V 分别是数据点的横坐标向量和纵坐标向量，它们有对等的地位。 常规最小二乘法认为纵坐标中有噪声而横坐标中无噪声显然是不合理的。而总体最小二乘法同时考虑 T 和 V 中的噪声，令代表横坐标 T 向量中的误差，则有 $(T-\varphi)\cdot a_{s,n}=V-\varepsilon$。在总体最小二乘法中，$a_{s,n}$ 的选取

应使得总体误差 $\Gamma=[\Delta\varepsilon]$ 的 Frobenius 范数最小，即 $\min\|\Gamma\|_F^2$，其中，

$$\|\Gamma\|_F^2=Tr\left[\Gamma^T\cdot\Gamma\right] \tag{10}$$

(10) 式中 $Tr[\cdot]$ 代表矩阵的迹。

基于总体最小二乘法的直线拟合可以归结为一个简单的二阶矩阵的特征向量求解，运算量很小[5]。该算法同时考虑了数据点在横坐标方向和纵坐标方向的不确定性，因而较常规的最小二乘法直线拟合具有更高的拟合精度。

因此，本文的数据流趋势分析算法决定采用总体最小二乘法对各个数据分段进行直线拟合，以提高趋势分析的精度。

2.2 可变滑动窗口算法

SW 算法的滑动窗口最大长度没有限制，当检测点阀值比较大时，窗口的长度可能很长使得趋势分析的误差变大；而 OSD 算法限定了最小滑动窗口长度，使得在最小滑动窗口内的突变点无法检测。

针对 SW 和 OSD 算法中滑动窗口存在的缺陷，本文提出一种动态改变设定窗口长度的可变滑动窗口算法，以合理的对数据序列进行分割。该算法首先设置数据段基准窗口长度和最长数据窗口长度，从当前数据段的起始点开始，对每一新到达的数据流元素都重新建立回归模型，提高精度。在当前数据段小于基准窗口长度时，用模型的拟合均方差与噪音函数 G 的返回值比较，检测基准窗口内是否存在异常点；在前数据段长度若大于等于基准窗口长度时，而且模型的拟合均方差大于预先设置的标准分割点阈值，则认为新到达的数据为当前数据段的分割点。如果当前数据段长度大于最长数据窗口长度，则回溯，从当前数据段的起点开始，从中寻找一个拟合均方差与标准分割点阀值最接近的数据点作为当前数据段的分割点。

可变滑动窗口算法解决 SW 算法和 OSD 算法的固定窗口问题，实现了对数据流的合理分割，因此趋势分析的精度得到了提高。具体算法的伪代码如下：

基准窗口长度：一般根据所应用领域分析的数据流的波形变化特点决定

最长数据窗口长度 MaxLength_Window：一般根据所分析的数据流波形变化的特点决定

标准分割点阀值：一般为所应用领域专家经验决定

噪音函数：一个接口函数，其功能为：当前数据段小于基准窗口长度时候调用，返回值为分割点阀值，该值要根据所应用领域的工艺条件决定

设模型的拟合均方差为 $\mu_{s,n}$，总体最小二乘法为 $f(\cdot)$

```
Begin
While (Y 不为空)  //由可（1）可知，Y 为所分析的数//据流
{ μ_{s,n} = f(Y_{s,n})  //由前面的分析可知，Y_{s,n} 为当前已接//收数据序列，n 为该序列的长度
 If (n < L)
   {If (μ_{s,n} > G(·))
      { v_{t_c} 是数据段 Y_{s,n} 的分割点，Y_{s,n} 的趋势特征值为
        β = [a_{s,n}, b_{s,n}]^T，并将后续到达的数据归入新
        的当前数据段，启动新的趋势分析过程}
     Else
       继续分析下一个到达的数据
   }
   If (n >= L)
     {If (μ_{s,n} > θ)
       {v_{t_c} 是数据段 Y_{s,n} 的分割点，Y_{s,n} 的趋势特征值为
        β = [a_{s,n}, b_{s,n}]^T，并将后续到达的数据归入新的
        当前数据段，启动新的趋势分析过程  }
      Else if (n<MaxLength_Window)
       {继续分析下一个到达的数据}
         Else
       {While (Y_{s,n} 不为空)
         {
            搜索拟合均方差方差与 θ 最为接近的
            数据点作为数据段 Y_{s,n} 的分割点，并将后
            续到达的数据归入新的当前数据段，启
            动新的趋势分析过程
          }   }   }   }
End
```

3 实验与分析

铝电解槽电压的有效趋势提取可提供槽况恶化的早期预警、评估槽况、工况的状态，对氧化铝浓度的判断提供支持信息。目前，贵铝槽控系统的数据采集中心每时每刻都会得到下位机传送上来的大量数据，槽电压受到阳极动作、效应发生、出铝等的影响变化较大，使得电压数据元素的值是不确定的，但存储单个数据结构体的大小是一致的，所以称得上是一种有序平稳的数据流。

论文基于总体最小二乘法建立回归模型以拟合各个分段的数据，采用可变滑动窗口算法进行分割点检测，对铝电解过程的重要参数(槽电压)进行实时趋势分析。该方法在贵铝二分厂电解 3 系列 3325 号电解槽上进行实时测试，取 2009-3-12 00:00 到 2009-3- 12 23:52 电解槽电压生产数据共 6370 个采样点作为测试样例，每两个电压采样点的时间间隔为 10 秒。本研究的程序

开发基于Microsoft Windows 2000 平台，采用C++ Builder 6.0 开发工具，实时数据库为Oracle 10g。

实验分别采用本文算法、SW 算法和 OSD 算法对上述数据流进行趋势分析，比较如下性能指标：1) 数据流分割的合理性；2)在不同阀值下趋势分析的精度。为了便于比较，对上述 3 种算法作如下限定：1)采用同一数据流进行分析；2) 分割点检测阀值相同。在满足上述限定下，对各算法的参数尽可能优化设置。其中，本文算法所涉及的参数进行如下考虑：结合铝电解过程 电压变化的特点(比如 电压受到阳极动作影响下波动较大)，综合考虑电解槽过程运行的工况、槽况，我们设置：基准窗口长度=T/4(T 为电压曲线波形的周期，该周期随运行工况可变，初始值为 110 分钟)，最长数据窗口长度 MaxLength_Window=*k(k 为可变参数，1.5<=k<=2，初始值为 1.5)，标准分割点阀值=r(r ∈ N，2<=r<=10)。

噪音函数需要根据计算的工况、槽况自适应的调整，以便能够准确的进行槽电压分割点的判断，使得槽电压的实时趋势分析更加的合理，也更能符合槽电压变化的特点，使得分析出来的趋势更能反映真实的电压趋势变化。当工况、槽况恶化时，电压曲线的波形变化常常表现为短时间内波动激烈，此时返回的阀值较小，以便找到窗口内的突变点。

三种算法的趋势分析效果如图 1 所示，取前 800 个电压采样点的三种算法的分割点显示对比如图 2 所示。

为了比较 SW 算法、OSD 算法和文中算法的趋势分析精度，还计算了三种算法的拟合均方误差。拟合均方误差 (MSE) 的计算公式为：$E_{MS}=\frac{1}{n}\sum_{i=1}^{n}\left[(at_i+b)-v_i\right]$，其中，n 表示数据段的长度，即电压采样点的点数；a 为该数据段的拟合的模型参数值；t_i 为实际电压，为横坐标时间。 表 1 和表 2 列出了在不同分割点检测阀值下三种算法的分割点点数 N_1、压缩比=6370/ N_1 及拟合均方误差 E_{MS}，图 3、图 4 显示了它们的压缩比和拟合均方误差比较。

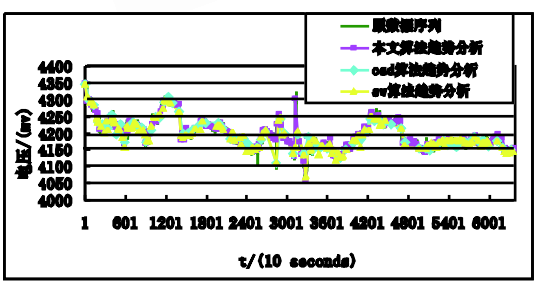

图 1　三种算法的数据流趋势分析

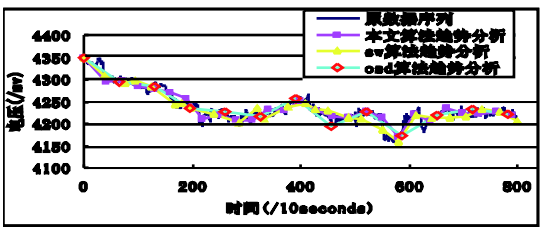

图 2　三种算法的数据流分割点比较

由图 1 可见，与 SW 算法和OSD 算法比较，本文的算法可以更好的逼近原始槽电压数据，由表 1、表 2、图 3、图 4 可见，文中算法与 SW 算法、OSD 算法的压缩比接近，但具有更高的逼近原始数据的精度，特别是当分割点检测的阀值较大的时候，相对 SW 算法、OSD 算法的精度改善更为明显。

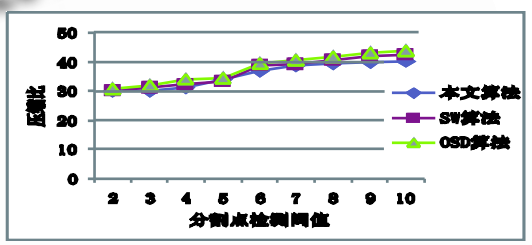

图 3　三种算法的压缩比比较

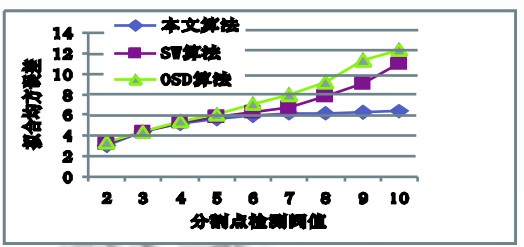

图 4　三种算法的拟合均方误差比较

表 1　三种算法的分割点数、压缩比、拟合均方误差比较(一)

分割点检测阀值		2	3	4	5	6
N_1	本文算法	213	212	205	198	172
	SW算法	210	205	187	190	164
	OSD算法	206	200	125	185	161
λ	本文算法	29.92	30.07	31.09	33.9	37.06
	SW算法	30.02	31.02	32.27	33.49	38.86
	OSD算法	30.8	31.8	33.99	34.43	39.38
E_{MS}	本文算法	3	4.33	5.13	5.56	6.92
	SW算法	3.21	4.35	5.26	5.80	6.26
	OSD算法	3.32	4.40	5.37	6.01	7.06

4　实验与分析

论文提出了一种新的数据流趋势分析方法，该方

表 2　三种算法的分割点数、压缩比、拟合均方误差比较(二)

分割点检测阈值		7	8	9	10
N_1	本文算法	165	161	160	158
	SW算法	163	157	151	150
	OSD算法	156	153	148	145
λ	本文算法	36.63	39.59	39.84	40.34
	SW算法	39.02	40.42	41.91	42.42
	OSD算法	40.62	41.62	42.99	43.88
E_{MS}	本文算法	6.14	6.15	6.27	6.35
	SW算法	6.67	7.85	9.02	11.03
	OSD算法	7.92	9.19	11.31	12.35

法采用总体最小二乘法对数据流进行分段拟合，提高了趋势分析精度；可变滑动窗口算法实现了对数据流的合理分割。该方法在贵铝二分厂电解 3 系列 3325 号电解槽上进行实时测试，测试结果表明，该算法准确分析预测了槽电压实时趋势变化，有效性较为明显。

参考文献

1 李岩,王惠文,叶明.数据流分析与技术研究.计算机工程与应用, 2008,44(15):8－9.

2 Melek WW, Lu Z, Kapps A, et al. Comparison of trend detection algorithms in the analysis of physiological time-series data. IEEE Trans. on Biomed Engineering, 2005,52(4):639－651.

3 张贤达.矩阵分析与应用.北京:清华大学出版社, 2004.

4 周黔,吴铁军.一种动态数据流的实时趋势分析算法.控制与决策,2008,(10):1183－1184.

5 杨云,孙群,朱长青.曲线数据压缩的总体最小二乘算法.西安电子科技大学学报(自然科学版), 2008,(5):947－948.