基于滑动多窗口的时间序列流趋势变化检测

李晓光,宋宝燕,张 昕
(辽宁大学信息学院,辽宁沈阳 110036)

摘 要: 趋势变化检测在时间序列流中有着非常广泛的应用.针对可变长的趋势变化检测问题,提出一种基于滑动多窗口的趋势变化检测方法,通过动态生成大尺度窗口,来适应可变长的趋势变化检测.针对内存约束下长趋势变化检测问题,提出一种基于增量 PLA 的窗口数据近似表示方法,给出了其欧式距离下的误差分析,进而提出一种误差修正方法来降低漏检率.大量实验表明,本文提出的检测方法具有高准确率且时间效率很高.

关键词: 时间序列流; 趋势变化监测; 分段线性逼近
中图分类号: TP311.13 **文献标识码:** A **文章编号:** 0372-2112 (2010) 02-0321-06

Sliding Multi-Windows Based Trend Change Detection on Time Series Stream

LI Xiao-guan, SONG Bao-yan, ZHANG Xin
(*School of Information, Liaoning University, Shenyang, Liaoning* 110036, *China*)

Abstract: Trend change detection has been applied widely to applications of time series stream. For the issue of detecting the change of variable length, a detection approach based on sliding multi-windows is proposed to scales up the sliding window to detect variable change. For the issue of long-term change detection with a memory constraint, a synopsis of incremental PLA is proposed to approximate to the original data, and then the error under the L_2 distance is analysis theoretically, by which an amendatory L_2 is given to reduce the ratio of true negative. The approach in this paper achieves quite high detection accuracy and efficiency within the extensive experiments.

Key words: time series stream; trend change detection; piecewise linear approximation(PLA)

1 引言

时间序列流的趋势变化是指某个时间段的时间子序列与参照时间子序列相比发生偏移或存在较大差异.与静态的时间序列相比,时间序列流是一种动态的时间序列,流中的元素是按时间顺序的、快速变化的、海量的和潜在无限的.由于时间序列流的这些特点,如何准确和快速地发现时间序列流上的变化是当前数据流研究的一个热点和难点.当前趋势变化检测的研究主要关注短期或长期趋势的变化.然而实际应用中,流的趋势变化是非常复杂的,通常具有长、短趋势变化共存的现象.本文重点研究了这种可变长的趋势变化检测方法.

时间序列流上的趋势变化主要通过比较当前窗口与某个历史窗口之间的距离来检测.主要的检测方法有两大类:一是采用多窗口技术[1],通过设定大小不同的多窗口来检测给定长度范围内的趋势变化,窗口内为某一时刻的固定数据.其类似于数据快照,通常假定待检测的趋势变化长度(窗口大小)满足内存限制,没有采用大纲方式存储数据,无法适用于流的长趋势变化检测.二是采用滑动窗口技术[2],通过比较连续相邻滑动窗口的距离来检测趋势变化,并采用 PAA 方式来存储数据.尽管采用了 PAA 大纲方式可以用于长趋势检测,但必须存储全部数据,且由于其为单窗口,只能对固定长度的趋势变化进行检测,且由于需要存储全部数据,只适用于时间序列上的变化检测.多窗口技术可以检测一定长度范围内的趋势变化,但范围必须事先指定且空间复杂度高.

针对与此,本文提出了一种基于滑动多窗口的趋势变化检测方法,该方法动态的生成不同长度的滑动窗口,并采用了分段线性逼近(Piecewise Linear Approximation,PLA)来近似表示窗口数据,通过计算窗口间的欧式距离判断是否发生变化.主要贡献点有:(1)提出一种

收稿日期:2008-11-17;修回日期:2009-09-21
基金项目:国家自然科学基金(No.60703068);国家自然科学基金(No.60873068)

基于滑动多窗口技术的检测算法,通过动态生成不同长度的滑动窗口,检测可变长的趋势变化;(2)提出一种基于 PLA 的窗口数据近似方法,详细分析了基于 PLA 的欧式距离误差,并提出一种误差修正方法;(3)提出一种窗口数据的 PLA 增量拟合方法,用于在给定窗口大小限制下,基于当前窗口的 PLA 增量生成大尺度 PLA 表示.

2 相关工作

文献[1]中提出的采用多窗口技术和概率分布相似度量来检测给定长度范围内的趋势变化.由于没有采用大纲方式构建窗口内数据的摘要,无法用于长趋势变化的检测.另外,窗口大小为事先指定且固定长度,无法调整窗口大小以适应趋势变化的长度变化.文献[2]提出通过比较连续相邻滑动窗口的距离来检测趋势变化,采用分段累积近似法 PAA 来近似存储窗口数据.PAA 方法误差比较大,并且为非增量拟合的,仅适用于静态时间序列数据.另外,在数据挖掘领域中也有大量的变化检测的方法.如基于统计的方法[3]、基于距离的方法[4]、基于密度的方法[5]、基于聚类的方法[6]及神经网络的方法[7]等.然而这些研究工作主要面向非时间序列流的,数据为无序的且必须存储全部数据才能进行检测.

由于流数据的无限性,流上变化分析通常需要构建大纲.目前常用的大纲除了前面所述的 PAA 还有 PLA[8]、DWT[10]、DFT[9]、Sketch[11]等.DWT 和 DFT 将时间序列由时域转为频域,适用于静态的时间序列分析.文献[12]提出一种可以用于时间序列流的 DWT 方法 incDWT,但在片断相似度计算时仍需重建,时间复杂度较高.而 Sketch 方法中对给定长度 l 的片段,如果提高概率,sketch 很大,且需为每一个定长时间序列计算 sketch,无法处理增量数据,只适用于静态时间序列.分段线性逼近法(PLA)为时间序列线性表示方法的线性回归方法,而线性内插是时间序列线性表示的另一种方法.就逼近的精确度而言,由于 PLA 采用距离误差最佳拟合,尤其是按照欧氏距离计算时,精确度相对于其他方法要高[8].

3 基于滑动多窗口的可变长趋势变化检测

3.1 相关概念与技术

定义 1 时间序列流:时间序列流为一序列 $S=(x_0,x_1,\cdots,x_n)$,其中 n 随时间增加,$\forall x_t\in S$,x_t 为一实数,n 为流的当前长度.

定义 2 子序列:给定一个长度为 n 的时间序列 S,长度为 m 的子序列 C 是时间序列 S 的一个片断,它是由 S 中长度为 $m\leqslant n$ 个连续数据组成.

定义 3 滑动窗口:令流 S 上的子序列 $SW_i=(x_i,\cdots,x_{i+W-1})$,$i+W-1\leqslant n$,则称 SW_i 为时刻 i 长度为 W 的滑动窗口.

定义 4 长度为 W 的趋势变化:给定流 S 上两个窗口大小为 W 的滑动窗口 SW_b 和 $SW_i(i\geqslant b+W)$,如果 $dist(SW_b,SW_i)>\xi$,则称时刻 i 流 S 发生长度为 W 的趋势变化,其中 $dist$ 是距离函数,ξ 是阈值.此时称 SW_b 为参照窗口,SW_i 为比较窗口.

本文中的距离函数 $dist$ 为欧式距离,即当给定两个长度为 n 的时间序列 S 和 Q,那么它们的欧氏距离 $dist(S,Q)$为:

$$dist(S,Q)=\sqrt{\sum_{i=1}^{n}(s_i-q_i)^2} \tag{1}$$

对于时间序列 $S(s_1,s_2,\cdots,s_n)$,其 PLA 表示为 $S(s'_1,s'_2,\cdots,s'_n)$其中$S'_t=at+b(t\in[1,n])$

$$a=\frac{12\sum_{t=1}^{n}(t-\frac{n+1}{2})s}{n(n+1)(n-1)} \tag{2}$$

$$b=\frac{6\sum_{t=1}^{n}(t-\frac{2n+1}{3})s_t}{n(1-n)} \tag{3}$$

那么对于 S,只需保存 PLA 系数 a 和 b 即可.

定义 5 时间序列 S 的长度为 l 的 PLA 表示.对于时间序列 $S(s_1,s_2,\cdots,s_n)$,设 S 分为 m 个长度为 l 的连续个子序列,即 $n=l\cdot m$,那么令 a_i 和 b_i 分别为第 i 个子序列的 PLA 表示的系数,则 S 的长度为 l 的 PLA 表示为 $S_{PLA}=<a_1,b_1;a_2,b_2;\cdots;a_m,b_m>$

给定两个时间序列 $S(s_1,s_2,\cdots,s_n)$和 $Q(q_1,q_2,\cdots,q_n)$,令 $S_{PLA}=<a_{11},b_{11};\cdots;a_{1m},b_{1m}>$和 $Q_{PLA}=<a_{21},b_{21};\cdots;a_{2m},b_{2m}>$分别为 S,Q 的长度为 l 的 PLA 表示,那么 S_{PLA}和Q_{PLA}之间的 PLA 距离 $dist_{PLA}(S,Q)$为:

$$dist_{PLA}(S,Q)=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{l}((a_{1i}-a_{2i})\cdot j+(b_{1i}-b_{2i}))^2} \tag{4}$$

令 $x_i=s_i-q_i$,则式(1)和式(4)分别可以表示为

$$dist^2(S,Q)=\sum_{i=1}^{m}\sum_{j=(i-1)\cdot l+1}^{i\cdot l}x_j^2 \tag{5}$$

$$\begin{aligned}dist^2_{PLA}(S,Q)&=\sum_{i=1}^{m}\Big(\frac{l(l+1)(2l+1)}{6}\cdot(a_{1i}-a_{2i})^2+l\cdot(l+1)(a_{1i}-a_{2i})\cdot(b_{1i}-b_{2i})+l\cdot(b_{1i}-b_{2i})^2\Big)\\&=\sum_{t=1}^{m}\Big(\frac{12}{l(l-1)(l+1)}\Big(\sum_{j=(i-1)\cdot l+1}^{i\cdot l}(j-(i-1)\cdot l)\cdot x_j\Big)^2\Big)-\frac{12}{l(l-1)}\Big(\sum_{j=(i-1)\cdot l+1}^{i\cdot l}(j-(i-1)\cdot l)\cdot x_j\Big)\cdot\sum_{j=(i-1)\cdot l+1}^{i\cdot l}x_j+\frac{2(2l+1)}{l(l-1)}\cdot\Big(\sum_{j=(i-1)\cdot l+1}^{i\cdot l}x_j\Big)^2\Big)\end{aligned} \tag{6}$$

3.2 变化检测算法

本文中变化检测方法针对不同长度的趋势变化,

采用动态生成参照窗口来检测,即在给定的时间范围内,如果未检测到变化,则在当前参照窗口的基础上,生成尺度更大的窗口.当窗口大小超过给定最大窗口限制时,则对新生成窗口中的数据采用 PLA 近似表示.具体的检测算法见算法 1.算法 1 中的窗口数据结构有两种:当窗口大小没有超过最大限制时,以数据点的形式保存;否则对窗口数据进行 PLA 近似且仅保存系数 a 和 b.算法中的 Σ^1 和 Σ^2 分别为 $\sum_{t=1}^{l} ts_t$ 和 $\sum_{t=1}^{l} s_t$,其中 s_t 是序列 t 时刻数据点,Σ^1 和 Σ^2 的更新和计算见 3.4 节.|…|表示窗口的大小,l 为 PLA 长度.

算法 1　SMWCDetect

输入:时间序列流 S,最小窗口的宽度 W,最大窗口 $W_{\max}$,时间范围阈值 α 和距离 ζ

输出:变化报告

(1) $curWin = 1$;
(2) append the consecutive stream data of size $2W$ to SW'_1 and SW_1; $SW_1.l = SW'1.l = 1$
(3) for each new arrival of data x
(4) for $i = 1 \cdots curWin$
(5) if($|SW_i|$ $SW_i.l < W_{max}$)
(6) append x to SW_i and remove the data of SW_i;
(7) if($dist(SW'_i, SW_i) > \zeta$)then report the change; $SW'_i \leftarrow SW_i$; append the next $|SW_i|$ to SW_i
(8) else
(9) wait for next $SW_i.l$ data N after x, and compute the Σ^1 and Σ^2 of N;
(10) append the (Σ^1, Σ^2) to SW_i, and remove the latest (Σ^1, Σ^2);
(11) if($dist_{PLA}(SW'_i, SW_i) > \zeta$) then report the change; $SW'_i \leftarrow SW_i$; wait for the next $|SW_i|$ $SW_i.l$ data and create the PLAs of length $SWi.l$ for SWi
(12) if there has been not change reported by SW_{curWin} for α timestamps, then
(13) if $|SW_{curWin}|$ $SW_{curWin}.l + \Delta \leqslant W_{\max}$ then
(14) create two new windows $SW'_{curWin+1}$ and $SW_{curWin+1}$ of size $|SW_{curWin}| + \Delta$;
(15) $SW'_{curWin+1}.l = SW_{curWin+1}.l = 1$;
(16) append the data of SW_{curWin} and the next Δ data to $SW'_{curWin+1}$;
(17) append the next $|SW_{curWin+1}|$ $SW_{curWin+1}.l$ data to $SW_{curWin+1}$;
(18) else
(19) create two new windows $SW'_{curWin+1}$ and $SW_{curWin+1}$ of size $W_{\max}$;
(20) $SW'_{curWin+1}.l = SW_{curWin+1}.l = S_{WcurWin}.l * 2$;
(21) append the data of SW_{curWin} and the next $SW_{curWin+1}.l * (|W_{\max}| - |SW_{curWin}|/2)$ data to $SW'_{curWin+1}$, and create the PLAs of length $SW'_{curWin+1}.l$ for $SW'_{curWin+1}$
(22) append the next $|W_{\max}|$ $SW_{curWin+1}.l$ data to $SW_{curWin+1}$ and create the PLAs of length $S_{WcurWin}.l$ for $SW_{curWin+1}$
(23) $curWin$ ++;

3.3　误差估计与修正

给定 S 和 Q,由于 $dist_{PLA}(S,Q) \leqslant dist(S,Q)$[8],则当 l 较大时会导致大量的漏检情况.令距离误差 $\varepsilon = dist(S,Q) - dist_{PLA}(S,Q)$,本文提出利用 ε 的期望值 $E\varepsilon$ 来修正 $dist_{PLA}$,即 $dist'_{PLA}(S,Q) = dist_{PLA}(S,Q) + E\varepsilon$.根据切贝谢夫不等式,有 $P(|\varepsilon - E\varepsilon| \leqslant \omega) \geqslant 1 - D\varepsilon/\omega$,那么给定一个 ω,以概率(至少为 $1 - D\varepsilon/\omega$)保证修正值 $E\varepsilon$ 与真实值 ε 接近(不超过 ω).给定 $S(s_1, s_2, \cdots, s_n)$ 和 $Q(q_1, q_2, \cdots, q_n)$,令 $x_i = s_i - q_i$,$C_{\max}$ 和 $C_{\min}$ 为 x_i 中的最大值和最小值,$C'_{\max}$, C'_{min} 分别为 x_i^2 的最大值和最小值.由于在时间序列流中,我们并不清楚即将到来具体的值,但可以设定一个最小值和最大值,并认为 x_i 取该区间任何一个值的可能性相同,则 x_i 符合 $[C_{\max}, C_{\min}]$ 区间的均匀分布,其期望和方差分别为:

$$E(x_i) = \frac{C_{\max} + C_{\min}}{2}, D(x_i) = \frac{(C_{\max} - C_{\min})^2}{12} \quad (7)$$

那么,对于给定 i,则 ix_i 也为随机变量,则有

$$E(i \cdot x_i) = \frac{i \cdot (C_{\max} + C_{\min})}{2}, D(i \cdot x_i) = \frac{i^2 \cdot (C_{\max} - C_{\min})^2}{12} \quad (8)$$

定理 1　当 l 足够大时,我们有:

$$\sum i \cdot x_i \sim N\left(\mu_1 = \frac{l(l+1)(C_{\max} + C_{\min})}{4}\right),$$

$$\delta_1^2 = B_N^2 = \sum D(i \cdot x_i) = \frac{(C_{\max} - C_{\min})^2}{12} \cdot \frac{l(l+1)(2l+1)}{6} \quad (9)$$

证明　根据李雅诺普夫定理,令 $\delta = 2$ 且令

$$ix_i = X_i \lim_{l\to\infty}\frac{1}{B_n^4}\sum E(X_i - EX_i)^4$$
$$= \lim_{l\to\infty}\frac{\sum(X_i - \frac{i(C_{\max}+C_{\min})}{2})^4\cdot\frac{1}{i(C_{\max}-C_{\min})}}{\frac{1}{72^2}(C_{\max}-C_{\min})^4\cdot l^2\cdot(l+1)^2(2l+1)^2} \quad (10)$$

由于

$$\lim_{l\to\infty}\frac{\sum(-\frac{i(C_{\max}+C_{\min})}{2})^4\cdot\frac{1}{i(C_{\max}-C_{\min})}}{\frac{1}{72^2}(C_{\max}-C_{\min})^4\cdot l^2\cdot(l+1)^2(2l+1)^2} \leqslant (10) \leqslant$$
$$\lim_{l\to\infty}\frac{\sum(iC_{\max}-\frac{i(C_{\max}+C_{\min})}{2})^4\cdot\frac{1}{i(C_{\max}-C_{\min})}}{\frac{1}{72^2}(C_{\max}-C_{\min})^4\cdot l^2\cdot(l+1)^2(2l+1)^2}$$
$$\lim_{l\to\infty}\frac{\sum(-\frac{i(C_{\max}+C_{\min})}{2})^4\cdot\frac{1}{i(C_{\max}-C_{\min})}}{\frac{1}{72^2}(C_{\max}-C_{\min})^4\cdot l^2\cdot(l+1)^2(2l+1)^2}$$
$$= \lim_{l\to\infty}\frac{\frac{1}{16}\sum i^3}{\frac{1}{72^2}(C_{\max}-C_{\min})\cdot l^2\cdot(l+1)^2(2l+1)^2}$$
$$= \lim_{l\to\infty}\frac{\frac{1}{16}\cdot\frac{1}{4}\cdot l^2(l+1)^2}{\frac{1}{72^2}(C_{\max}-C_{\min})\cdot l^2\cdot(l+1)^2(2l+1)^2}$$

当 $l\to\infty$ 时,

$$\lim_{l\to\infty}\frac{\sum(-\frac{i(C_{\max}+C_{\min})}{2})^4\cdot\frac{1}{i(C_{\max}-C_{\min})}}{\frac{1}{72^2}(C_{\max}-C_{\min})^4\cdot l^2\cdot(l+1)^2(2l+1)^2} = 0.$$

同理可证不等式右边 =0,所以当 $l\to\infty$ 时,式(10)等于0,则 $\lim P(\frac{1}{B_N}\sum(ix_i - E(ix_i)) \leqslant x) \sim N(0,1)$ (11)

又 $\lim P(\frac{1}{B_N}\sum(ix_i - E(ix_i)) \leqslant x)$

$$= \lim P(\frac{1}{B_N}(\sum ix_i - \frac{l(l+1)}{2}\cdot\frac{C_{\max}+C_{\min}}{2})) \leqslant x)$$

所以结合(11)可得

$$\sum ix_i \sim N(\frac{l(l+1)(C_{\max}+C_{\min})}{4}, B_n^2).$$ 证毕.

定理 2 当 l 足够大时,我们有:

$$\sum x_i \sim N(\mu_2 = \frac{l(C_{\max}+C_{\min})}{2}, \delta_2^2 = \frac{l(C_{\max}-C_{\min})^2}{12}) \quad (12)$$
$$\sum x_i^2 \sim N(\mu_3 = \frac{l(C'_{\max}+C'_{\min})}{2}, \delta_3^2 = \frac{l(C'_{\max}-C'_{\min})^2}{12}) \quad (13)$$

根据中心极限定理,容易证明定理 2.

定理 3 给定长度为 l 的时间序列 S 和 Q,令 $C_{\max}$ 和 $C_{\min}$ 为 x_i 中的最大值和最小值,$C'_{\max}$、$C'_{\min}$ 分别为 x_i^2 的最大值和最小值,$\varepsilon = dist(S,Q) - dist_{PLA}(S,Q)$,则有

$$E\varepsilon = \frac{l}{2}(C'_{\max}+C'_{\min}) - \frac{2l+3}{6}(C_{\max}^2+C_{\min}^2) \quad (14)$$

证明 对于式(6),有:令

$$A = \sum ix_i,\ B = \sum x_i,\ C = \sum x_i^2$$ 则,

$$dist_{\mathrm{PLA}} = \frac{12}{l(l-1)(l+1)}A^2 - \frac{12}{l(l-1)}AB + \frac{2(2l+1)}{l(l-1)}B^2$$
$$= \frac{12}{l(l-1)}(\frac{A}{\sqrt{l+1}} - \frac{\sqrt{l+1}}{2}B)^2 + \frac{1}{l}B^2 \quad (15)$$

由于 $\varepsilon = dist\text{-}dist_{\mathrm{PLA}}$,则

$$E\varepsilon = E(dist\text{-}dist_{\mathrm{PLA}})$$
$$= EC - \frac{12}{l(l-1)}E(\frac{A}{\sqrt{l+1}} - \frac{\sqrt{l+1}}{2}B)^2 - \frac{1}{l}EB^2 \quad (16)$$

其中,由式(13)可得

$$EC = \mu_3 \quad (17)$$

利用方差的性质 $D(x) = Ex^2 - (Ex)^2$,结合定理 1 和定理 2,分别可得

$$E(\frac{A}{\sqrt{l+1}} - \frac{\sqrt{l+1}}{2}B)^2$$
$$= D(\frac{A}{\sqrt{l+1}} - \frac{\sqrt{l+1}}{2}B) + [E(\frac{A}{\sqrt{l+1}} - \frac{\sqrt{l+1}}{2}B)]^2$$
$$= \frac{1}{l+1}DA - \frac{l+1}{4}DB + [\frac{1}{\sqrt{l+1}}EA - \frac{\sqrt{l+1}}{2}EB]^2$$
$$= \frac{\delta_1^2}{l+1} - \frac{l+1}{4}\delta_2^2 + (\frac{\mu_1}{\sqrt{l+1}} - \frac{\sqrt{l+1}}{2}\mu_2)^2 \quad (18)$$
$$EB^2 = DB + (EB)^2 = \delta_2^2 + \mu_2^2 \quad (19)$$

把式(17)、(18)、(19)分别代入式(16)可得,

$$E\varepsilon = \mu_3 - \frac{12}{l(l-1)}(\frac{\delta_1^2}{l+1} - \frac{l+1}{4}\delta_2^2)$$
$$+ (\frac{\mu_1}{\sqrt{l+1}} - \frac{\sqrt{l+1}}{2}\mu_2)^2) - \frac{\delta_2^2+\mu_2^2}{l} \quad (20)$$

把式(9)、(12)、(13)代入式(20)可得

$$E\varepsilon = \frac{l}{2}(C'_{\max}+C'_{\min}) - \frac{2l+3}{6}(C_{\max}^2+C_{\min}^2)$$ 证毕

3.4 PLA 增量拟合

本文提出一种增量拟合 PLA 方法,无需存储全部原始数据,只需在长度为 l 的分段系数基础上,直接获得长度为 $2l$ 的 PLA 分段系数.

定理 4 给定流 $S=(s_1,s_2,\cdots,s_n)$,设流 S 中自时刻 i 长度为 l 的两个相邻的子序列 $C_i=(s_i,s_{i+1},\cdots,s_{i+l-1})$ 和 $C_{i+l}=(s_{i+l},s_{i+l+1},\cdots,s_{i+2l-1})$,令 $\sum_i^1 \equiv \sum_{t=1}^{l} tS_{i+t-1}$, $\sum_i^2 \equiv \sum_{t=1}^{l} S_{i+t-1}$, $\sum_{i+l}^1 \equiv \sum_{t=1}^{l} tS_{i+l+t-1}$, $\sum_{i+l}^2 \equiv \sum_{t=1}^{l} S_{i+l+t-1}$,则对于由 C_i 和 C_{i+l} 构成的长度

为 $2l$ 的子序列 $C=(s_i,s_{i+1},\cdots,s_{i+2l-1})$来说,令 $\sum^1\equiv\sum_{t=1}^{2l}tS_{i+t-1}$, $\sum^2\equiv\sum_{t=1}^{2l}S_{i+t-1}$,则

$$\sum^1=\sum_i^1+\sum_{i+l}^1+l\sum_{i+l}^2 \tag{21}$$

$$\sum^2=\sum_i^2+\sum_{i+l}^2 \tag{22}$$

证明　对于式(21)有,

$$\begin{aligned}\sum^1&\equiv\sum_{t=1}^{2l}tS_{i+t-1}=\sum_{t=1}^{l}tS_{i+t-1}+\sum_{t=1}^{l}(t+l)S_{i+l+t-1}\\&=\sum_{t=1}^{l}tS_{i+t-1}+\sum_{t=1}^{l}tS_{i+l+t-1}+l\sum_{t=1}^{l}S_{i+l+t-1}\\&=\sum_i^1+\sum_{i+l}^1+l\sum_{i+l}^2\end{aligned}$$

对于式(22)有,

$$\begin{aligned}\sum^2&\equiv\sum_{t=1}^{2l}S_{i+t-1}=\sum_{t=1}^{l}S_{i+t-1}+\sum_{t=1}^{l}S_{i+l+t-1}\\&=\sum_i^2+\sum_{i+l}^2.\end{aligned}$$

证毕.

根据式(21)和式(22),可以根据相邻的长度为 l 的子序列的 $\sum_{t=1}^{l}tS_t$ 和 $\sum_{t=1}^{l}S_t$ 计算长度为 $2l$ 的子序列的 $\sum_{t=1}^{2l}tS_t$ 和 $\sum_{t=1}^{2l}S_t$.进一步,根据 PLA 系数计算式(2)和式(3)可得 a 和 b 分别为 $\sum_{t=1}^{l}tS_t$ 和 $\sum_{t=1}^{l}S_t$ 的函数.那么,我们只要为每个 PLA 片段保存其 $\sum_{t=1}^{l}tS_t$ 和 $\sum_{t=1}^{l}S_t$ 值,即可计算出其 a 和 b.

4　实验分析

4.1　实验设置

本节主要验证了 SMWCDetect 算法中所采用的主要技术,包括距离修正的验证、检测质量的验证和运行效率的验证.实验选取文献[1]中趋势检测算法作为参照比较算法,其具体描述见相关工作.实验采用人工生成数据和实际时间序列数据作为测试数据集,其中人工生成数据是周期变化的正弦函数.实际数据为太阳黑子数据[13].

4.2　实验结果及分析

4.2.1　距离修正　图 1 和图 2 为太阳黑子数据上距离修正比较结果,其中 *dist* 为原始数据的欧式距离(式(1)),$dist_{PLA}$为基于 PLA 分段的欧式距离(式(6)),$dist_{PLA}+E\varepsilon$ 为经过修正后的欧式距离.$E\varepsilon$ 的计算公式见式(14),为了剔除噪声数据对参数估计的影响,这里 $C_{\min}$ 和 $C_{\max}$取值为以双边置信度 10% 下的原始数据差值的最小和最大值.从图 1 可以看出同 $dist_{PLA}$相比,$dist_{PLA}+E\varepsilon$ 更为接近原始欧式距离 *dist*.图 2 为原始数据的欧式距离与修正后的欧式距离差值($dist-dist_{PLA}+E\varepsilon$)分布图,可以看出,92.8% 的差值大于零,平均相对差值(($dist-dist_{PLA}+E\varepsilon$)/($dist-dist_{PLA}$))为 0.57,且其中 56.3% 的相对差值小于 1%.那么,经过修正后欧式距离仅能产生最多为 7.2% 的误检率,但是可减少最多为 92.8% 的漏检率.

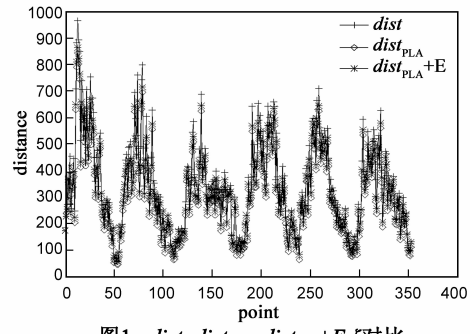


图1　*dist*, $dist_{PLA}$, $dist_{PLA}+E$ ξ对比

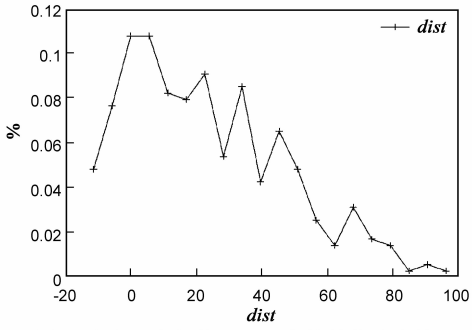


图2　真实距离与修正后距离的差值分布情况

4.2.2　效率分析　由于没有与本文采取策略类似的算法,因此本文主要与在原始数据上的检测算法(naiveDetect)进行比较,naiveDetect 与 SMWCDetect 主要不同在于前者假设窗口足够大,无须进行 PLA 近似.图 3 为在不同 PLA 分段长度 l 和最大窗口 $w_{\max}$下 naiveDetect 与 SMWCDetect 平均运行时间的比较结果,由于 naiveDetect 不近似表示数据并假设窗口足够大,因此其运行时间只与当前窗口数据大小有关且随着窗口大小增加呈线性增长,而 SMWCDetect 的运行时间与 l 和 $w_{\max}$有关.给定 $w_{\max}$,当数据量小于 $w_{\max}$其运行时间与 naiveDetect 类似为线性增长,而当大于 $w_{\max}$时,SMWCDetect 需要进行 PLA 增量拟合,根据公式 2、3、21 和 22,其 PLA 增量拟合时间复杂度为常量阶的,而欧式距离计算的复杂度为 $O(w_{\max})$,因此 SMWCDetect 的时间复杂度为常量阶.那么给定 l,SMWCDetect 的时间复杂度为 $O(w_{\max})$,图 3 也验证了上述分析.

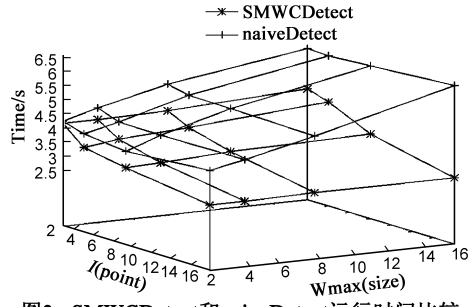


图3　SMWCDetect和naiveDetect运行时间比较

4.2.3 趋势变化检测算法比较 本节对 SMWCDetect 与文献[1]中 FIND-CHANGE 算法在检测准确率上进行比较,SMWCDetect 与 FIND-CHANGE 的初始窗口设置一致,SMWCDetect 窗口增量为 4,最大窗口为 16,距离修正中采取以双边置信度 10%下的原始数据差值的最小和最大值来估计 C_{min} 和 C_{max}. 图 4 为部分实验结果,其中 SMWCDetect-1 为基于距离修正的检测方法,SMWCDetect-2 为无距离修正的检测方法. SMWCDetect-1 的误检率为 45.94%,漏检率为 8.76%,SMWCDetect-2 的误检率为 38.32%,漏检率为 12.68%,而 FIND-CHANGE 的误检率为 69.82%. 由于数据中存在大于最大窗口长度的变化,而 FIND-CHANGE 为固定多窗口,因此导致 FIND-CHANGE 无法检测,如图中 110 时刻. 而 SMWCDetect 采用可变多窗口,因此可以很好适用于可变变化的检测,其平均漏检率较 FIND-CHANGE 降低 12.61%. 而由于采用距离修正,SMWCDetect-1 的漏检率较 SMWCDetect-2 降低 3.82%,而误检率增加了 764%. 在运行效率上,尽管 SMWCDetect 动态生成窗口,但由于采用 PLA 近似以及 PLA 增量拟合方法,SMWCDetect 的运行时间优于 FIND-CHANGE 算法,其中在 10M 数据量下,SMWCDetect 算法运行时间为 0.832s,而 FIND-CHANGE 算法为 3.233s.

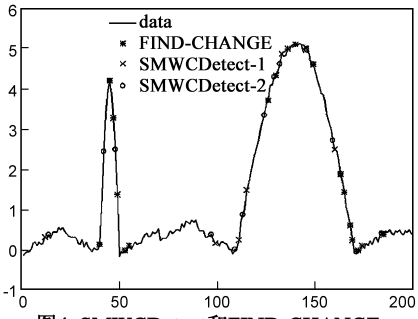


图4 SMWCDetect和FIND-CHANGE算法检测趋势变化

5 结论

针对可变长的趋势变化检测问题,本文提出了一种基于动态生成的滑动多窗口的变化检测方法 SMWCDetect. SMWCDetect 通过动态生成大尺度窗口,适应可变长的趋势变化检测. 针对内存约束下长趋势变化检测问题,提出一种基于增量 PLA 的窗口数据近似表示方法,给出了欧式距离下误差分析,进而提出一种误差修正方法来降低漏检率. 大量实验表明,本文提出的检测方法具有高准确率且时间效率很高,特别是对于可变趋势变化检测有较低的漏检率.

参考文献:

[1] Shai Ben David, Johannes Gehrke, Daniel Kifer. Detecting change in data stream[A]. In Proc. of VLDB[C]. San Francisco: Morgan Kaufmann, 2004. 180 – 191.

[2] Eamonn Keogh, Jessica Lin, Ada Fu. HOT SAX: Finding Most Unusual Time Series Subsequence: Algorithms and Applications [J]. Knowledge and Information Systems, 2007, 11(1): 1 – 27.

[3] Yamanishi K, Takeuchi J. Discovering Outlier Filtering Rules from Unlabeled Data: Combining a Supervised Learner with an Unsupervised Learner[A]. In Proc. of KDD [C]. New York: ACM, 2001. 389 – 394.

[4] Knorr E M. Outliers and Data Mining: Finding Exceptions in Data[D]. Canada: Univesity of British Columbia, 2002.

[5] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying Density based Local Outliers[A]. In Proc. of SIGMOD[C]. New York: ACM, 2000. 427 – 438.

[6] He Zengyou, Xu Xiaofei, Deng Shengchun. Discovering Cluster based Local Outliers[J]. Pattern Recognition Letters, 2003, 24 (9/10): 1651 – 1660.

[7] Harkins S, He H, Willams G J, et al. Outlier Detection Using Replicator Neural Networks[A]. In Proc. of the 4th International Conference on Data Warehousing and Knowledge Discovery [C]. Springer-Verlag : London, 2002. 170 – 180.

[8] QiuXia Chen, Lei Chen, Xiang Lian, Yunhao Liu , Jeffrey Xu Yu. Indexable PLA for Efficient Similarity Search [A]. In Proc. of VLDB[C]. New York: ACM, 2007. 435 – 446.

[9] Y L Wu, D Agrawal, A E Abbadi. A comparison of DFT and DWT based similarity search in time-series databases[A]. In Proc. of CIKM[C]. New York: ACM, 2000. 488 – 495.

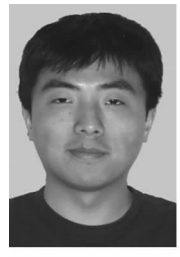
[10] K P Chan, A W Fu. Efficient time series matching by wavelets [A]. In Proc. of ICDE [C]. Washington: IEEE Computer Society, 1999. 126.

[11] Piotr Indyk, Nick Koudas, S. Muthukrishnan. Representative trends in massive time series data sets using sketches[A]. In Proc. of VLDB [C], San Fransisco: Morgan Kaufmann, 2000: 363 – 372.

[12] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin J. Strauss. One-Pass Wavelet Decompositions of Data Streams [J]. IEEE Transactions on knowledge and data engineering. 2003, 15(3): 541 – 554.

[13] Sun data [OL], www.schoolsobservatory.org, 2007.

作者简介:



李晓光 男,1973 年生于辽宁省沈阳市,博士,副教授. 主要研究领域为 XML 数据库、数据挖掘、信息检索、流数据分析.
E-mail: xgli@lnu.edu.cn

宋宝燕 女,1965 年生于辽宁省沈阳市,博士,教授,副院长. 主要研究领域为 RFID、XML 数据库、流数据分析.
E-mail: bysong@lnu.edu.cn

张 昕 男,1979 年生于内蒙古赤峰,辽宁大学信息学院讲师,博士,主要研究方向为数据库、复杂网络与信息融合.