基于 Lip-切比雪夫方法的数据流异常检测分析①

Lip-Chebyshev-Based Method for Data Stream Anomaly Detection and Analysis

贺 忠 (广州科技贸易职业学院 计算机系 广东 广州 511442)

摘 要: 近年来计算机软硬件性能飞速发展，数据流被广泛应用于传感器阵列、互联网信息的传输、数据决策支持系统(DSS)等诸多领域。数据流的传输速度越来越快，传输规模越来越大，人们对数据处理的时间要求越来越高。利用 Lipschitz 指数和切比雪夫定理，提出了一种数据流异常检测的算法，与传统方法相比具有同时包含整体和局部数据信息、计算量稳定的特点。

关键词: 数据流 异常检测 Lipschitz 指数 实时性 小波分析

近年来，数据流在工业控制、商业管理、市场分析和工程设计等生产经营领域有着广泛的应用。随着人们对数据流的分析处理研究不断深入，各种新的处理手段应运而生。抽样检测、动态网格、分形技术、小波变换等方法都可以应用到数据流的处理之中。但在众多的处理手段中，利用基于小波概要结构的方案吸引了众多学者的目光，由此产生了许多处理数据流的算法。通过把信息从时域变换到小波域，利用小波分解把输入数据变换成一个小波系数序列，而后通过小波域的处理方法(如 SPC 和小波的时间尺度图)[1,2]能够获取信号的奇异性信息。

现有的方法大多是仅从数据整体角度[3-5]或仅从局部角度来检测数据流。从整体的角度分析数据能充分利用从已知数据获得的知识，在没有先验知识的情况下利用已知数据来获得知识是十分重要的，但这样就忽视了数据在时域的相关性。新到的数据与之前的若干数据有很高的局部相似性，通过当前的局部信息可以很好的预测将来的数据，这种预测可以来进行异常检测。可是单纯的从局部信息来预测，既无法充分利用先验知识也无法从已知数据中获取知识。将来的新数据也是整体数据的一部分，仅从有限的局部判断整体也是不科学的，因此理想的方法应该既能体现整体知识又能充分利用数据的时域局部相关性。切比雪夫定理可以提供数据流的整体知识，通过小波系数局部模极大值计算的 Lipschitz 指数能很好的指示信号局部奇异性信息。本文以 Lipschitz 指数和切比雪夫定理为工具，提出了一种能实时检测出数据流中异常情况的有效实时检测算法。

1 数据流和异常

数据流是一种动态的数据，与传统数据相比，它具有动态和无法事先预知的特性。由于无法预知数据流的规模，不可能把整个数据完全的存储下来。数据流的处理和接收必须是同时进行的，即进行实时的响应和实时的处理。如果判断一个新数据点是否异常所需的计算规模与数据流的规模相关，那么计算的强度会越来越大，最终导致实时性的要求无法保证。因此数据流异常的实时检测应该保证每一个新来数据所需的计算规模是恒定的。

计算数据的均值和标准差所需的计算量与数据规模无关。已知 n 个数据的均值和标准差，新到一个数据后，需要计算 n+1 个数据的均值和标准差时，可以采用下面的公式:

$$\overline{x_{n+1}} = \frac{\left(n * \overline{x_n} + x_{n+1}\right)}{n+1} \tag{1}$$

$$\sigma_{n+1} = \sqrt{\frac{\left(\sigma_n^2 + \overline{x_n}^2\right) * n + x_{n+1}^2}{n+1} - \overline{x_{n+1}}^2} \tag{2}$$

① 收稿时间:2009-04-07

利用窗口模式，挖掘窗口内数据的信息可以获得充分的局部信息，并同时保证了计算量不会随着时间的推移而逐渐增大，从而保证了实时性的要求。

异常是一个抽象概念，数据流的应用也很广泛，很难找到一个定义或通过一个表达式来概括工程中可能出现的所有异常情况。我们把数据流异常分为三种典型情况，分别为奇异点，突变，和概念漂移。其他的异常可以通过这三种典型异常线性表出。这样的划分可以确保通过研究典型异常的检测来应对工程中出现的绝大多数异常。

异常点是指远离分布整体的量测值。产生异常点的原因很多，可能是过失差错，可能是样本点没有落在实验设计的范围之内，也可能就是极少数就来自此分布的奇异点。从数学角度来说异常点就是一种奇异点，从工程上看异常点就是指信号中的突变点。突变是指某一序列在给定时间点前某一时段的平均值与时间点后另一时段的平均值之间的差异具有充分的统计显著性，并且均值最终会回归到之前的水平。该给定的时间点就是突变点。从时间序列中发现定性知识会面临概念漂移问题。随着时间的流逝，数据发生变化会导致以往的概念不准确甚至失效。概念漂移与突变不同，概念飘移可以是一个长期的过程，同时概念漂移之后一般情况下不会回到漂移前的状态。从数学上看，概念漂移在波形上表现为台阶的形状。

假设正常的情况下的信号是一个值为常数 0 的时间序列。图 1 自上而下分别是该信号在 0 时刻出现异常点、突变和概念漂移异常的用 Matlab 绘制的示意图。

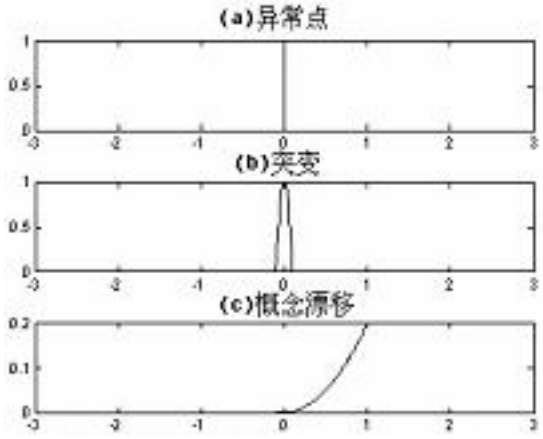

图 1　异常点、突变和概念漂移的数学波形表示

2　切比雪夫定理和Lipschitz指数

描述一个信号序列有很多种方法，使用平均值，中位数等统计参数都能用来估计信号总体的大小，而方差和标准差则可以用来衡量信号的总体稳定程度。这些统计特性参量计算十分简单，并且有着良好的统计特性，能客观的反应信号的特征，因此在信号的处理中有着很重要的地位。对于任何一个数据集合及不小于 1 的实常数 h，至少有$1-\frac{1}{h^2}$比例的样本位于以总体均值为中心、h 倍总体标准差为半径的区间之内，这就是切比雪夫定理。切比雪夫定理从整体角度对信号的分布进行了约束。

一般来说，函数的奇异性可以用它的可微性来表示。若函数在某处间断或某阶导数不连续，则称该函数在此处有奇异性。在数学上用 Lipschitz 指数(简称 Lip 指数)描述函数在某点的奇异性。

Lip 指数α定义如下[6]：对于函数 f(x)，设有非负整数 n，　$n\leqslant\alpha\leqslant n+1$，如果存在常数 A>0 以及 n 次多项式，$p_n(x)$

$$\left|f(x)-p_n(x-x_0)\right|\leq A\left|x-x_0\right|^{\alpha} \qquad (3)$$

对于 $x\in(x_0-\delta,x_0+\delta)$成立，则称 f(x)在点 x_0 是 Lip α 的，Lip α 指数表明了 f(x)与 n 次多项式比较，光滑程度的大小。

信号 Lip 指数的大小可以用小波变换来估计[7]，为了考察不同尺度小波变换与信号奇异性的关系，这里采用小波变换的卷积形式，把尺度 s 作为自变量来对待，此时小波变换的表达式如下：

$$W_f(s,x)=\int_R f(t)\psi(\frac{x-t}{s})dt=f*\psi_S(t) \qquad (4)$$

由于小波基ψ具有紧支撑特性，对于式(4)，当

$$\left|\frac{x-t}{s}\right|>k,\quad \psi(\frac{x-t}{s})=0$$

因而式(4)在区间[x-sk,x+sk]之外为 0，即：

$$W_f(s,x)=\int_{x-sk}^{x+sk}\frac{1}{s}f(x)\psi(\frac{x-t}{s})dt \qquad (5)$$

当 s→0 时，小波变换就反映了信号在 x 点的局部性态，即可以利用小波变换来判断函数的局部奇异性。

利用小波变换系数的局部极大值求出奇异点比较容易，常用这种方法检测奇异点[8]。将小波变换的各尺度的模极大值相连，可以得到小波系数的模极大值线。模极大值线上的小波模值和点 x_0 的 Lip 指数存在如下关系：

$$W_f(s,t) \le As^{\alpha} \tag{6}$$

式中 A 为正整数。当尺度 s→0 时，极大值线 x→x_0。如果找到极大值线，根据其模值变化率和时间坐标的渐近性可确定相应奇异点 x_0 的位置与 Lip 指数 α。对于二进离散小波变换，当尺度 s 以二进变化时，不存在模极大值线，因而需要确定极值在相邻尺度上的传播。如果把尺度 j 上某一点的模极大值记作 α_j，$\alpha_j = |W_{2^j} f(x_0)|$，那么在各尺度相应位置处的模极大值可构成序列$\{\alpha_j\}$，在 j 较小时可以做一定的近似：

$$\alpha_j = |W_{2^j} f(x_0)| \cong A2^{j\alpha} \tag{7}$$

$$\alpha_{j+1} = |W_{2^j} f(x_0)| \cong A2^{j(\alpha+1)} \tag{8}$$

将(7)(8)两式取对数相减可以得到：$\alpha = \log_2 \alpha_{j+1} - \log_2 \alpha_j$。当 j 的最大值为 6 时，则可以用

$$\alpha = \frac{(\log_2 \alpha_6 - \log_2 \alpha_3) + (\log_2 \alpha_5 - \log_2 \alpha_2) + (\log_2 \alpha_4 - \log_2 \alpha_1)}{9}$$

等取平均值的方法来得到较准确的 Lip 指数。Lip 指数计算的 Matlab 实现已有相应的研究[9]，这里不再赘述。

我们构造了一个含 600 个数据点的实验性数据流，它包含了异常点，突变和概念漂移三种情况。图 2 是用 Matlab 实现的对该时间序列进行高斯小波变换和相应的时间尺度图与小波模极大曲线，图中蓝色曲线是小波系数的局部模极大值曲线：

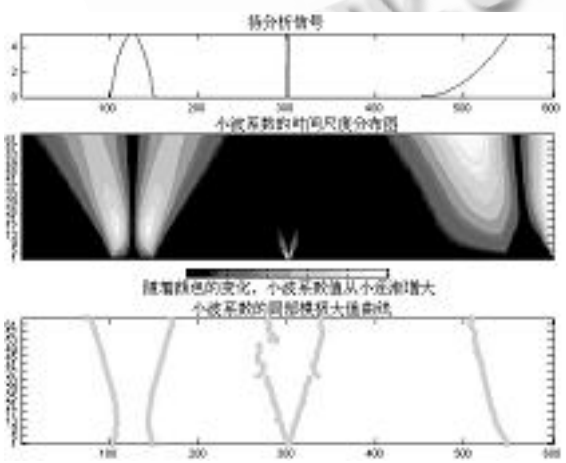

图 2　对一组人造数据的仿真

在图 2 中，各奇异点处 Lip 指数都出现了明显的变化，出现极值。通过时间尺度图与小波模极大曲线也可以较为准确地判断出异常出现的大致位置。由此可见，Lip 指数是指示信号奇异性、进行异常检测的有力工具。

3　Lip-切比雪夫算法

Lip-切比雪夫算法是一个滑动窗口算法。它的基础是利用 Lipschitz 指数和切比雪夫定理这两个工具来检测异常点、突变和概念漂移这三类典型异常。只需要计算总体均值，总体标准差，窗口内均值，窗口内标准差(分别用 $\bar{x}$，σ，$\overline{x_{WIN}}$ 和 σ_{WIN} 表示)以及和 Lip 指数就可以实现这个算法。

下面我们具体分析三种典型异常的检测方法。对于异常点检测，我们可以事先设定一个阈值，例如我们要保证对异常点检测的准确性在 90%以上，按照切比雪夫定理，均值 3.1 个方差的范围内会包容 90%以上的数据。根据这个事实我们可以确定下面的异常点检测算法：

① 将新到的数据点移入当前窗口。

② 计算更新总体均值，总体标准差，窗口内的均值，窗口内标准差四个参数。

③ 检查新数据点是否同时在($\bar{x}$ -3.1，$\bar{x}$ +3.1)和($\overline{x_{WIN}}$ -3.1 σ，$\overline{x_{WIN}}$ +3.1 σ)之外，如果是则认为新数据是一个异常点，否则就不是异常点。

异常点是最简单的一类异常情况，判断方法也比较简单。也可以用窗口的 Lip 指数来检测异常点，但会比上面的方法要复杂。

突变是一种剧烈的异常情况，它包含了一个陡峭的上升期和下降期，因此出现突变时，窗口内的切比雪夫定理无法满足。突变会使数据急剧变化从而产生奇异性，所以窗口内由小波变换系数模极大曲线指示的 Lip 指数会出现显著变化。根据本文第二部分对切比雪夫定理的描述，当 h=1 时，比例 $1-\frac{1}{h^2}$ 的值为 0，所以数据在总体均值附近一个标准差范围内时，可以把它理解为数据的正常波动(根据具体情况，允许的数据的波动范围可以由用户自由确定)。为了排除正常的数据波动，我们应该利用总体均值和总体标准差检验数据是否在均值的一个标准差范围内。具体算法如下：

① 将新到的数据点移入当前窗口。

② 计算更新总体均值，总体标准差，窗口内的均值，窗口内标准差四个参数及窗口内的 Lip 指数。

③ 检查窗口内切比雪夫定理是否满足(这里 h 的取值应根据实际情况，可以与检测异常点时不同)，若满足则进行其他检查(可能是概念漂移、正常的数据波动等情况)，否则检查窗口 Lip 指数是否变化显著。

④ 若 Lip 变化显著则检查数据是否在正常的数据波动的允许范围内，若不满足则认为出现了突变的情况。

与突变相比较，概念漂移是一个缓和与长期的过程。概念漂移的初期和数据的正常波动很容易混淆。当概念漂移发生时，窗口内的数据总是满足切比雪夫定理的。但由于数据的变化，仍会导致数据产生奇异性，因此窗口内的 Lip 指数会产生显著的变化。为了防止正常的数据波动被误判为概念漂移，和突变检测的方法类似，我们还要利用总体均值和总体标准差进行进一步检验排除数据的正常波动。具体算法如下：

① 将新到的数据点移入当前窗口。

② 计算更新总体均值，总体标准差，窗口内的均值，窗口内标准差四个参数及窗口内的 Lip 指数。

③ 检查窗口内切比雪夫定理是否满足(这里 h 的取值应根据实际情况，可以与检测异常点时不同)，若不满足则进行其他检查(可能是突变等情况)，否则检查窗口 Lip 指数是否变化显著。

④ 若 Lip 变化显著则检查数据是否在正常的数据波动的允许范围内，若不满足则认为出现了概念漂移的情况。

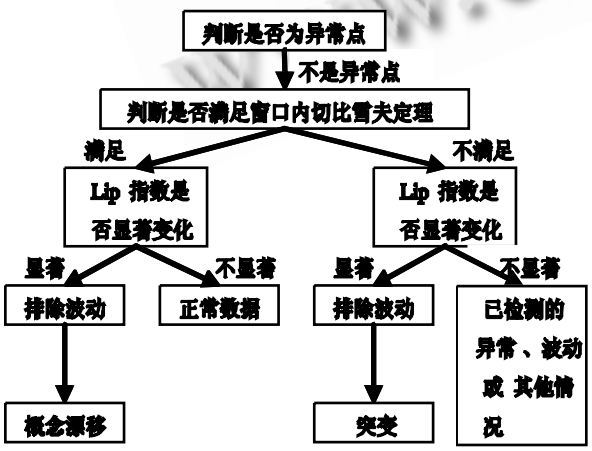

图 3　判断次序示意图

实际应用中，不需要依次检查新数据是否为异常点、突变和概念漂移，这会使算法变的复杂。算法只要进行三到四次判断就可以分辨出数据是否异常及属于哪种异常。首先判断新数据是否为异常点，当不是异常点时，判断窗口内切比雪夫定理是否满足并检验 Lip 指数的变化情况。排除数据正常波动的情况后，就可以确定数据是否出现了异常。我们把判断的流程以树的方式在图 3 中给出。

4　算法效果分析

Karras和Mamoulis的单次扫描小波概要结构的最大误差率算法是从整体角度进行检测的一个典型算法，它的时间和空间复杂度都与数据的规模成接近线性的关系。SPC 和阈值时间尺度图是从局部角度进行检测的典型算法。SPC 方案的一个明显缺点是在整体漂移的情况下，它的平均运行长度(ARL)不好。阈值时间尺度图方法则随着数据规模的增大而性能变强。这些算法除了有在引言中谈到了缺点外，计算的规模会随着数据的规模和数据本身的特点而有所改变。对于特殊的数据算法的性能会变得不好。

Lip-切比雪夫算法则一方面克服了单纯从整体角度或局部角度来检测数据流的弊端，而且没有发现 Lip—切比雪夫算法的性能与数据的特点有直接关系；另一方面假设窗口大小为 W，数据的规模为 N，那么算法的总体时间复杂度和 W*N 成线性关系，空间复杂度与 W 成线性关系，性能相比之前的算法是稳定的。由于窗口大小在算法运行前就确定了，此时 W 可看作是一个常数。对每个新来的数据点来说计算量与 W 成正比，也是恒定的，保证了实时性要求。另外通过改变算法中窗口长度、高斯函数和小波分解的最大尺度等参数可以改变算法的性能从而适应不同场合应用的需要。

我们用 ECG、Space Shuttle、Surveillance 三个数据集测试了使用 C 语言编写的 Lip-切比雪夫算法，结果找到了三个数据集中大部分的异常点，对异常点的定位也较准确。

5　总结

数据流的分析处理具有重要的现实意义，在日常的生产生活中有着很多的应用，如心电图 QRS 波检测和疾病暴发的早期检测等。把 Lip-切比雪夫算法应用到具体的生产生活中去，需要根据具体情况，结合应

(下转第 5 页)

(上接第 64 页)

用本身的特点，来调节算法中设置的参数，从而达到最佳的检测效果。Lip—切比雪夫算法的时间和空间复杂度均与数据流的整体规模无关，而只和窗口的大小等用户设置的参数相关，保证了实时性的要求。

参考文献

1 Jeong MK, Lu JC, Wang N. Wavelet-Based SPC Procedure for Complicated Functional Data. Interna- tional Journal of Production Research. 2006, 44(4):729 – 744.

2 Jeong MK, Chen D, Lu JC. Thresholded Scalogram and Its Application in Process Fault Detection. Applied Stochastic Models in Business and Industry. 2003, 19(3):231 – 244.

3 Karras P, Mamoulis N. One-pass wavelet synopses for maximum-error metrics. Proc. of the 31st international conference on Very large data bases. August 30-September 02, 2005, Trondheim, Norway.

4 Garofalakis M, Gibbons PB. Wavelet synopses with error guarantees. Proc. of SIGMOD Conf, 2002:476 – 487.

5 Garofalakis M. Kumar A. Deterministic wavelet thresholding for maximum-error metrics. Proc. of PODS, 2004:166 – 176.

6 Mallat S, Hwang WL. Singularity Detection and Processing with wavelets. IEEE transac- tions on information theory, 1992,38(2):617 – 643.

7 Mallat S.杨力华,戴道清,黄文良,湛秋辉,译.信号处理的小波导引(第二版).北京:机械工业出版社, 2002:122 – 163.

8 林晖,张优云.运用小波分析处理结构优化问题.计算力学学报, 2000,17(3):278 – 286.

9 龙兴明,周静.连续小波变换的一维信号检测.重庆邮电学院学报(自然科学版), 2004,16(3):77 – 80.