·先进计算与数据处理· 文章编号:1000-3428(2018)05-0014-05 文献标志码:A 中图分类号:TP301.6

基于多级 Haar 小波变换与 KS 统计的突变点快速探测方法

宋巧红,齐金鹏,张 煜

(东华大学 信息科学与技术学院,上海 201620)

摘 要: 结合多级 Haar 小波变换与 KS 统计理论,提出一种对时序数据突变点的快速探测方法(HWKS),对标准参考序列以及待检测序列分别构建均值二叉搜索树和差值二叉搜索树。基于改进的 KS 检验方法给出二叉树搜索的 2 种策略,进而构建实现时序数据突变点快速检测的 HWKS 理论框架。运用 HWKS 对模拟的时序数据进行检测,与 HW 方法、T 方法和 KS 方法的比较结果表明,该方法在对时序数据的突变点进行检测时的误差较小、用时最短、准确度较高。

关键词: KS 统计理论;多级 Haar 小波变换;二叉搜索树;时序数据;突变点检测

中文引用格式: 宋巧红,齐金鹏,张 煜. 基于多级 Haar 小波变换与 KS 统计的突变点快速探测方法[J]. 计算机工程,2018,44(5):14-18,24.

英文引用格式: SONG Qiaohong, QI Jinpeng, ZHANG Yu. Fast Abrupt-point Detection Method Based on Multistage Haar Wavelet Transform and KS Statistic[J]. Computer Engineering,2018,44(5):14-18,24.

Fast Abrupt-point Detection Method Based on Multistage Haar Wavelet Transform and KS Statistic

SONG Qiaohong, QI Jinpeng, ZHANG Yu

(College of Information Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 Combined with multi-level Haar wavelet transform and KS statistic theory, this paper proposes a fast detection method for time series data abrupt-point, that names as HWKS. The mean binary search tree and the difference binary search tree are constructed respectively for the standard reference sequence and the sequence to be detected. Based on the improved KS test method, two methods of binary tree search are proposed, and the HWKS theory framework for rapid detection of time series data mutation point is realized. HWKS is used to detect the simulated time series data. Comparison results on HW method, T method and KS method show that HWKS has less error, shorter time and higher accuracy when detecting time series data abrupt-point.

【Key words】 KS statistic theory; multistage Haar wavelet transform; binary search tree; time series data; abrupt-point detection

DOI: 10.19678/j.issn.1000-3428.0046777

0 概述

在数据挖掘和统计领域,时序数据突变点的检测已经引起了广泛的关注[1-2],大部分方法是通过比较时间序列样本在过去和现在的概率分布来检测的[3],这种根据分布的不同来检测出异常点的方法灵活性较差[4-6]。在统计方面,已经探索出了一些用于突变点检测的方法,比如 KS 统计理论[7]。KS 统计理论是一种非参数的统计理论,其量化了样本的经验分布函数和参考分布的累积分布函数之间的距离,尤其是两样本的 KS 检验方法,被广泛用于比较 2 个样本,因为它对 2 个样本的经验分布函数的位置和形状参数的差异特别敏感[8]。另一方面,小波变换对于异常点的检测有很好的前景。小波变换可以很容易地从不同的时间或空间距离上提取数据的分布特征[9]。小波分析的核心是多分辨率分析,可以把其中一个信号分解成不同大小的分辨率级别的子信号。小波的定位、正交性和多速率滤波等特性是对信号平稳性和瞬态性分析必不可少的条件[10]。

然而,这些方法大多很耗时,且由于时间复杂度的问题并不适合处理分析大数据[11]。此外这些方法对于一些无效的波动不是很敏感,尤其是 2 个端

基金项目: 国家自然科学基金(61305081,61104154);上海市自然科学基金(16ZR1401300,16ZR1401200)。

作者简介: 宋巧红(1990—),女,硕士研究生,主研方向为大数据异常检测;齐金鹏,副教授、博士;张 煜,硕士研究生。

收稿日期: 2017-04-14 **修回日期:** 2017-05-15 **E-mail:** songqh2016@163.com

点附近的突化。为了实现对时间序列检测的及时性[12-13],本文在多级 Harr 小波变换与 KS 统计理论的基础[14-15]上提出一种新的快速探测突变点的理论框架,简称 HWKS(Haar Wavelet and KS),并与 HW 方法、T 方法和 KS 方法从耗时、命中率、误差以及准确度 4 个方面进行比较。

1　HWKS 的理论框架

HWKS 的框架结构如图 1 所示。首先以正常的序列 X 作为参考序列,然后对其和待检测序列 Z 通过多级 Haar 小波变换,分别构建均值二叉搜索树(TcA)和差值二叉搜索树(TcD)。其次,用改进的 KS 统计理论对 TcA 中的根节点到叶子节点中的异常点进行快速检测。最后,对模拟的时序数列进行突变点检测,以验证该方法的有效性。

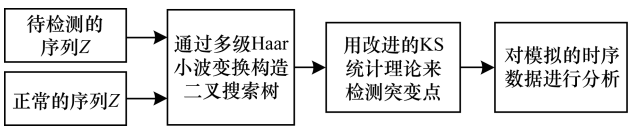

图 1　HWKS 突变点检测过程

1.1　均值二叉搜索树和差值二叉搜索树的构建

一般而言,可利用多级 Haar 小波变换方法,将离散的时序信号 $Z=\{z_1,z_2,\cdots,z_N\}$ 解析成不同的频域分量,并表示如下:

$$Z\xrightarrow{H_k}(cA^k\mid cD^k\mid cD^{k-1}\mid\cdots\mid cD^2\mid cD^1) \tag{1}$$

其中,cA 和 cD 分别为均值参数和差值参数。多分辨率分析(Multi-Resolution Analysis,MRA)是小波分析的核心[16],根据 MRA,可以概念化小波变换的过程,将总体的 N 分解成一小段一小段 n。向量 $\boldsymbol{v}_i$ 和 $\boldsymbol{w}_i$ 分别代表缩放信号和小波基向量。离散信号 Z 的均值信号 A^k 和差值信号 D^i 可以表示为:

$$Z=A^k+\sum_{i=1}^{k}D^i,1\leqslant k\leqslant \mathrm{lb}\ N \tag{2}$$

$$A^k=(Z\cdot V^k)V^k=\sum_{i=1}^{N/2^k}(Z\cdot v_i^k)v_i^k=\sum_{i=1}^{N/2^k}(cA_{k,i})v_i^k \tag{3}$$

$$D^i=(Z\cdot W^i)W^i=\sum_{j=1}^{N/2^i}(cD_{k,j})w_j^i \tag{4}$$

因此,可以得到如下方程:

$$Z=A^k+\sum_{i=1}^{k}D^i=\sum_{i=1}^{N/2^k}(Z\cdot v_i^k)v_i^k+\sum_{i=1}^{k}\sum_{j=1}^{N/2^i}(Z\cdot w_j^i)w_j^i=cA^k\cdot V^k+\sum_{j=1}^{k}cD^j\cdot W^j \tag{5}$$

$$A^k=cA^k\cdot V^k=(cA_{k,1},cA_{k,2},\cdots,cA_{k,N/2^k})\cdot(v_1^k,v_2^k,\cdots,v_{N/2^k}^k)=(a_1^k,a_2^k,\cdots,a_{N-1}^k,a_N^k) \tag{6}$$

$$D^k=cD^k\cdot W^k=(cD_{k,1},cD_{k,2},\cdots,cD_{k,N/2^k})\cdot(w_1^k,w_2^k,\cdots,w_{N/2^k}^k)=(d_1^k,d_2^k,\cdots,d_{N-1}^k,d_N^k) \tag{7}$$

在式(6)中,V_i^k 是多级 Haar 小波变换的参考信号。在式(7)中,W_j^k 是多级 Haar 变换的小波基信号,且 $|V_i^k|=|W_j^k|=N$。

因此,可将时序数据 Z 分解成多维的均值参数矩阵(***McA***)和差值参数矩阵(***McD***):

$$\boldsymbol{McA}=\begin{bmatrix}cA_{1,1}&\cdots&cA_{1,N/2}\\ \cdots&cA_{k,j}&0\\ cA_{m,1}&0&0\end{bmatrix}$$

$$\boldsymbol{McD}=\begin{bmatrix}cD_{1,1}&\cdots&cD_{1,N/2}\\ \cdots&cD_{k,j}&0\\ cD_{m,1}&0&0\end{bmatrix} \tag{8}$$

其中,$0\leqslant k\leqslant m=\mathrm{lb}N,1\leqslant j\leqslant N/2^k$。

综上,可将时序数据 ***Z*** 分解成多维的均值参数矩阵(***McA***)和差值参数矩阵(***McD***)。然后,将参数矩阵 ***McA*** 和 ***McD*** 分别映射到对应的均值二叉搜索树(TcA)和差值二叉搜索树(TcD)的各层子节点。分别通过 ***McA*** 和 ***McD*** 来构造 TcA 和 TcD 中的各级非叶子节点。同时,叶子节点直接来自 Z 中的元素。多级 Harr 小波变换将待检测的序列 Z 分解成 TcA 和 TcD 的过程,如图 2 所示。

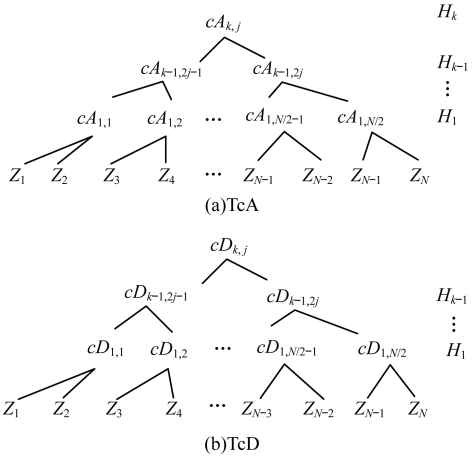

图 2　待检测序列的 Z 分解

1.2　基于改进 KS 统计理论的 HWKS

KS 统计是一种非常有用的非参数方法,被广泛应用于 2 个样本的比较。它对 2 个样本的经验分布函数的位置和形状参数的差异都特别敏感。假定一个时间序列 $Y=\{y_1,y_2,\cdots,y_N\}$,可以表示为 $Y=f(i/N)+X,\ i=1,2,\cdots,N$,其中 $X=\{x_i\}_{i=1,2,\cdots,N}$ 是独立同分布的随机变量,f 是未知分布的噪声信号。那么,就可以用分布函数 $F_m(x)$ 来表示正常的时间序列 X,同时用分布函数 $G_n(x)$ 来表示异常的时间序

列 Y。待检测的时间序列 Z 表示如下：

$$Z=\{X,Y\}=\{Z_1,Z_2\}=\{z_1,z_2,\cdots,z_c,z_{c+1},z_{c+1},\cdots,z_n\}$$

为了检测 Z 中的突变点，可以通过改进的 KS 统计理论来评估 X 分布和 Z 分布之间的距离，如式(9)所示。

$$D_{mn}(x)\underset{-}{\Delta}\left(\frac{mn}{m+n}\right)^{1/2}\sup_{x\in R}|G_n(x)-F_m(x)| \tag{9}$$

如果 Z 中有个突变点 c，那么就存在一个值 z_c，使得 $F_m(z_c)\neq G_n(z_c)$，其中，$D_{mn}(z_c)>\delta$，$z_c\in[z_1,z_n]$，$\delta\in\mathbb{R}$。然后，可以根据时间序列 X 和 Z 来分别表示它们的经验累积分布函数 $F_m(x)$ 和 $G_n(x)$，分布函数 $F_m(x)$ 和 $G_n(x)$ 用来统计低于 x 级样本点的比例。对于任何不动点 $x\in\mathbb{R}$，由大数定律可得：

$$F_m(x)=\frac{1}{m}\sum_{i=1}^{c}I(x_i\leqslant x)\rightarrow EI(x_i\leqslant x)=F(x) \tag{10}$$

$$G_n(x)=\frac{1}{n}\sum_{j=1}^{n}I(z_j\leqslant x)\rightarrow EI(z_j\leqslant x)=G(x) \tag{11}$$

可以表示 Z 中的第 j 个元素 z_j，同时在 HWKS 方法中定义一个新的元素 $z_{k,j}$：

$$z_j=a_{k,j}+\sum_{i=1}^{k}d_{i,j}\Rightarrow z_j\geqslant a_{k,j} \tag{12}$$

$$z_{k,j}=\frac{1}{2^k}\left(\sum_{i=a}^{b}z_i\right)=\frac{(\sqrt{2})\hat{}k}{2^k}cA_{k,j}=\frac{(\sqrt{2})\hat{}(k+1)}{2^k}a_{k,a}=\cdots=\frac{(\sqrt{2})\hat{}(k+1)}{2^k}a_{k,b}\leqslant z_a,\cdots,z_b \tag{13}$$

其中，$cA_{k,j}$ 是 TcA 中的非叶子节点，$z_{k,j}$ 是根据 $cA_{k,j}$ 新定义的元素，且 $1\leqslant j\leqslant N/2^k$，$a=2^k(j-1)+1$，$b=2^k\times j$。

可以为 HWKS 定义一个改进的 KS 统计理论方法：

$$D'_{mn}(k,j)\underset{-}{\Delta}\left(\frac{mn}{m+n}\right)^{1/2}\sup_{z_{k,j}\in R}|G_n(z_{k,j})-F_m(z_{k,j})|=\left(\frac{nm}{n+m}\right)^{1/2}\sup_{z_{k,j}\in R}\left|\left\{\frac{1}{n}\left(\sum_{j=1}^{n}I(z_j\leqslant z_{k,j})\right)-\frac{1}{m}\sum_{i=1}^{m}I(x_i\leqslant z_{k,j})\right\}\right| \tag{14}$$

在 TcA 中选定的节点 $cA_{k,j}$，用 $D'_{mn}(k,j)$ 来测试 X 和 Z 之间的分布距离，$D'_{mn}(k,j)$ 中最大值的位置说明 Z 中有可能产生突变。因此，可以定义 X 和 Z 中的 KS 检验为：

$$H_0:F_m=G_n \text{ vs. } H_0:F_m\neq G_n \tag{15}$$

为了检验 H_0，制定一个策略：

$$\delta=\begin{cases}H_0:D'_{mn}\leqslant c\\ H_1:D'_{mn}>c\end{cases} \tag{16}$$

假如 $\sup_k\sup_j|D'_{mn}(k,j)|\leqslant C(\alpha)$，其中，$\alpha$ 为置信区间。若 H_0 假设成立，那就说明没有突变点。另一方面，如果 $\sup_k\sup_j|D'_{mn}(k,j)|>C(\alpha)$，假设 H_1 成立，说明检测到突变点。

1.3 HWKS 框架的异常点搜索策略

为了检测时间序列 Z 的突变点，需要在 TcA 的根节点和叶子节点之间建立一条准确和快速的最优路径。第 1 个搜索策略如下所示：

策略 1 假定 TcA 中的非叶子节点是 $cA_{k,j}$，它左面的子节点和右面的子节点分别是 $cA_{k-1,2j-1}$ 和 $cA_{k-1,2j}$。

1)若($D'_{mn}(k-1,2j-1)>D'_{mn}(k-1,2j)$ 且 $D'_{mn}(k-1,2j-1)>C(\alpha)$)，则选择左面的子节点 $cA_{k-1,2j-1}$ 作为 TcA 中当前的搜索路径。

2)若($D'_{mn}(k-1,2j-1)<D'_{mn}(k-1,2j)$ 且 $D'_{mn}(k-1,2j)>C(\alpha)$)，则选择右面的子节点 $cA_{k-1,2j}$ 来作为 TcA 中当前的搜索路径。

如果满足 $(D'_{mn}(k-1,2j-1)>D'_{mn}(k-1,2j)$，表明 $cA_{k-1,2j-1}$ 所覆盖的左子树比 $cA_{k-1,2j}$ 所覆盖的右子树存在一个重要的距离分布，反之亦然。也就是说序列 Z 中突变点在左边发生的概率比右边发生的概率大。如果满足 $D'_{mn}(k-1,2j-1)>C(\alpha))$，意味着分布距离超过了一个相同的数据分布的临界值。策略 1 表明，如果序列 Z 中有个突变点，会选择左子树或者右子树中分布距离更大的一边来参与当前路径的搜索，未被选择的一边将被舍弃掉。因此，在 TcA 中，通过 lb(n)步可以得到一个从根节点到叶子节点的最优搜索路径。

然而，如果($D'_{mn}(k-1,2j-1)==D'_{mn}(k-1,2j)$或者($\max(D'_{mn}(k-1,2j-1),D'_{mn}(k-1,2j))<C(\alpha)$)，那么策略 1 对于突变点的检测是无效的。因此，在 TcD 的基础上构建一个新的搜索策略。

策略 2 若满足 $D'_{mn}(k-1,2j-1)==D'_{mn}(k-1,2j)$或者 ($\max(D'_{mn}(k-1,2j-1),D'_{mn}(k-1,2j))<C(\alpha)$)，则选择 TcD 中的非叶子节点 $cD_{k,j}$，其左侧的子节点和右侧的子节点分别为 $cD_{k-1,2j-1}$ 和 $cD_{k-1,2j}$。

1)如果满足 $|cD_{k-1,2j-1}|>|cD_{k-1,2j}|$，选择 TcA 中的左侧子节点 $cA_{k-1,2j-1}$ 作为当前的搜索路径。

2)如果满足 $|cD_{k-1,2j-1}|<|cD_{k-1,2j}|$，选择 TcA 中的右侧子节点 $cA_{k-1,2j}$ 作为当前的搜索路径。

根据上文所说的 2 种策略，HWKS 就是用这 2 种基本的算法来检测时间序列 Z 中的突变点。在策略 2 中，可以依据 TcA 中选定的非叶子节点 $cA_{k,j}$ 来计算 X 和 Z 之间的分布距离。在这个函数中，置信区间设置为 $\alpha=0.05$，$C(\alpha)=1.325\ 8$。为了简单

起见,只输出大于 $C(\alpha)$ 节点的值 D'_{mn},否则输出 0。根据上述 2 种搜索策略,可以得到从根节点到叶子节点的最优的搜索路径,然后就可以找到待检测时间序列 Z 中的突变点。

2　实验结果与分析

通过仿真的时序数据来评估 HWKS 的可行性,先对仿真数据进行一次检测,再对仿真数据进行多次检测。最后,设置不同的样本长度和突变点位置,对每组数据都进行 400 次重复实验,通过与 KS 方法、HW 方法和 T 方法的比较来分析 HWKS 方法的耗时、命中率、误差和准确度。

2.1　对数据的一次检测

图 3 为模拟的待检测时序数据。每个待检测的时序数列 Z 的长度都为 N,由 2 个部分组成,一部分为正常的时序数列,长度为 K,正常的序列服从均匀分布;另一部分为不正常的序列,长度为 NK,通过数据拼接的方式,将均匀分布替换为服从高斯分布的序列。图 3 中数据的长度为 32,突变点的位置为 6。

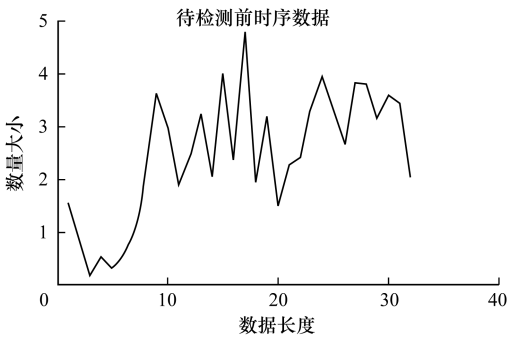

图 3　模拟的待检测时序数据

用这 4 种方法对图 3 的时序数据进行检测,检测结果如图 4 所示。为了更加清楚地查看检测结果,将图 4 的数据整理在表 1 中。

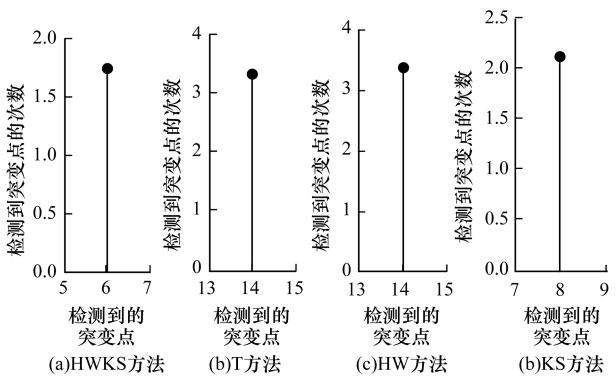

图 4　4 种方法的检测结果

表 1　一次检测的结果

方法	耗时/s	误差
HWKS 方法	0.035	0
KS 方法	0.002	8
HW 方法	0.081	8
T 方法	0.046	2

从表 1 可以看出,HWKS 的误差最小,耗时相对较小。KS 虽然耗时较小,但是误差很大。HW 耗时较长,T 方法的误差比 HWKS 方法大。所以,综合考虑,HWKS 在这 4 种方法中,效果最好。

2.2　对不同突变点不同数据长度的多次检测

先设置一个固定的突变点,待检测样本的长度 $N=128$,$K=111$。通过 40 次的仿真实验来检测突变点的位置。图 5(a) ~ 图 5(d)为对数据进行处理后的分布,图 5(e) ~ 图 5(h)为对检测到的不同位置的突变点的位置进行拟合,从图中可看出命中的突变点的大概位置。

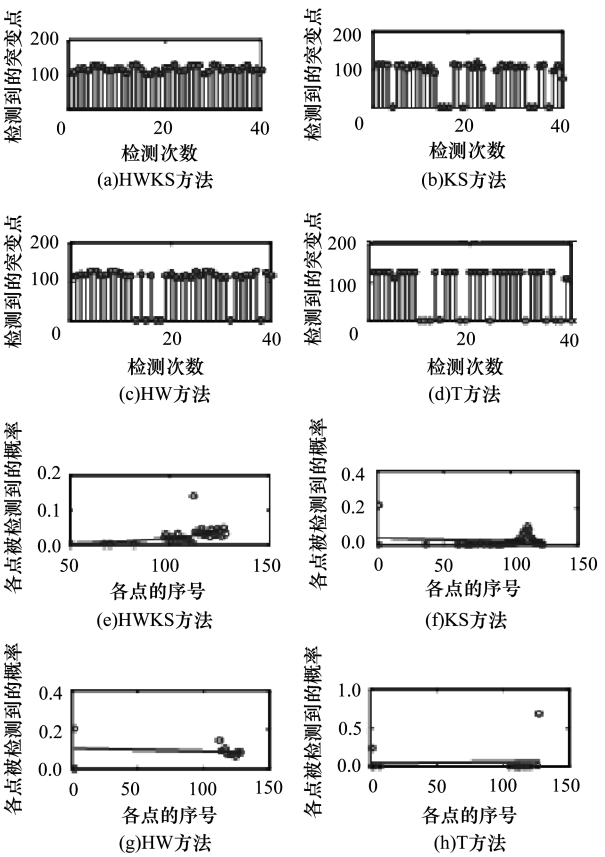

图 5　多次仿真的实验结果

为了更加清楚地查看检测结果,将数据整理在表 2 中。对同一个突变点进行多次检测时,可以发现 HWKS 方法的准确度最高,耗时最小。为了进

一步验证 HWKS 方法的可行性,设置了样本的不同长度,并设置不同的突变点位置,对检测的时间长短、命中率、误差和准确度进行检测。检测结果如表 3 所示。并将表 3 中各个指标的平均值整理在表 4 中。

表 2 多次仿真的实验数据

方法	耗时/s	命中率	误差	准确率
HWKS 方法	0.14	0.09	0	1.00
KS 方法	0.22	0.08	28	0.78
KS 方法	32.04	0.11	16	0.88
T 方法	2.86	0.002	17	0.87

表 3 多次检测结果的时间、命中率、误差、准确度

方法	评估内容	$N=2^4$, $k=3$	$N=2^5$, $k=5$	$N=2^6$, $k=7$	$N=2^7$, $k=111$	$N=2^8$, $k=189$	$N=2^9$, $k=447$	$N=2^{10}$, $k=789$	均值
HWKS	耗时/s	0.12	0.15	0.14	0.14	0.14	0.21	0.24	0.16
	命中率	0.19	0.18	0.10	0.09	0.05	0.02	0.15	0.11
	误差	4	7	15	0	22	16	28	13.14
	准确率	0.75	0.78	0.76	1.00	0.91	0.97	0.97	0.88
KS	耗时/s	0.03	0.04	0.10	0.22	0.55	1.91	6.70	1.36
	命中率	0.002	0.001	0.030	0.080	0.090	0.090	0.070	0.05
	误差	3	4	5	28	0	0	1	5.86
	准确率	0.81	0.88	0.92	0.78	1.00	1.00	0.99	0.91
HW	耗时/s	32.04	32.04	32.04	32.04	32.07	32.05	32.05	32.05
	命中率	0.12	0.09	0.04	0.11	0.03	0.05	0.00	0.06
	误差	4	11	28	16	33	30	112	33.43
	准确率	0.75	0.66	0.56	0.88	0.87	0.94	0.89	0.79
T	耗时/s	0.47	0.79	1.50	2.86	4.50	9.75	19.26	5.59
	命中率	0.030	0.030	0.030	0.002	0.001	0.010	0.020	0.020
	误差	3	8	17	17	17	71	67	28.57
	准确率	0.81	0.75	0.73	0.87	0.93	0.86	0.94	0.84

表 4 多次统计的平均值

方法	耗时/s	命中率	误差	准确率
HWKS 方法	0.16	0.11	13.14	0.88
KS 方法	1.36	0.05	5.86	0.91
HW 方法	32.05	0.06	33.43	0.79
T 方法	5.59	0.02	28.57	0.85

从统计的平均值可以看出 HWKS 方法的准确度相对较高,虽然 KS 的准确度较高,但是耗时比 HWKS 要大很多。当样本的尺寸较小,且突变点发生在左边界或者右边界时,HWKS 的命中率要比 KS 高。

相比于 HWKS 方法和 KS 方法,在检测突变点时,HW 方法需要耗费更多的时间,而且命中率不是很高。

表 3 的结果表明,HWKS 对于具有较小尺寸 N 的样本的左边界和右边界附近的较小显著数据波动具有更好的性能和灵敏度,由于计算时间较短,命中率较高,因此 HWKS 是用于模拟时间序列上的突变点检测的较好方法。

3 结束语

本文结合多级 Haar 小波变换与 KS 统计理论,给出对时序数据突变点的快速探测方法 HWKS。该方法主要利用多级 Haar 小波变换与 KS 统计理论,对标准参考序列以及待检测序列分别构建均值二叉搜索树(TcA)和差值二叉搜索树(TcD)。并基于改进的 KS 检验方法提出二叉树搜索的 2 种策略,进而完成对时序数据突变点的快速检测的 HWKS 理论框架的构建。最后,用模拟的时序数据进行验证,结果表明,与 HW 方法、T 方法和 KS 方法相比,HWKS 方法在对突变检测时的误差较小,用时最短,准确度较高。综合考虑,HWKS 方法是对突变点进行检测的一种有效的方法。

参考文献

[1] 李国杰. 大数据研究的科学价值[J]. 中国计算机学会通讯,2012,8(9):8-15.

[2] 蒋 涛,冯玉才,朱 虹,等. 时序数据挖掘概述[EB/OL]. [2017-04-14]. http://doc.mbalib.com/view/8f6ae3ed41ef4cd4ec9207ae75d1cbf8.html.

[3] 秦首科. 数据流上的异常检测[D]. 上海:复旦大学,2006.

[4] MANIKOPOULOS C, PAPAVASSILIOU S. Network intrusion and fault detection: a statistical anomaly approach[J]. IEEE Communications Magazine, 2002, 40(10):76-82.

(下转第 24 页)

率大幅提高的同时,能更好地保持甚至改善分类效果。

为了使所给算法能更有效地用于大规模数据集,特别是大数据处理,还需要对算法在效率上做进一步改进,这是下一步要做的一项研究内容。

参考文献

[1] VAPNIK V. The Nature of statistical learning theory[M]. New York,USA:Springer,1995.

[2] JOACHIMS T. Text categorization with support vector machine:learning with many relevant features[C]// Proceedings of the 10th European Conference on Machine Learning. New York,USA:ACM Press,1998:137-142.

[3] 王宪亮,吴志刚,杨金超,等. 基于 SVM 一对一分类的语种识别方法[J]. 清华大学学报(自然科学版),2013,53(6):808-812.

[4] 孙俊涛,张顺利,张 利. 基于联合支持向量机的目标跟踪算法[J]. 计算机工程,2017,43(3):266-270.

[5] DONG Jianxiong,KRZYZAK A,SUEN C Y. Fast SVM training algorithm with decomposition on very large data sets[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(4):603-618.

[6] LI Boyan, WANG Qiangwei, HU Jinglu. Fast SVM training using edge detection on very large datasets[J]. IEEE Transactions on Electrical and Electronic Engineering,2013,8(3):229-237.

[7] JUNG H G. Support vector number reduction: survey and experimental evaluations novel[J]. IEEE Transactions on Intelligent Transportation Systems,2014,5(2):463-476.

[8] 包文颖,胡清华,王长忠. 基于多粒度数据压缩的支持向量机[J]. 南京大学学报(自然科学版),2013,49(5):637-643.

[9] 焦李成,张 莉,周伟达. 支撑矢量预选取的中心距离比值法[J]. 电子学报,2001,29(3):383-386.

[10] 李红莲,王春花,袁保宗,等. 针对大规模训练集的支持向量机的学习策略[J]. 计算机学报,2004,27(5):715-719.

[11] PANDA N,CHANG E Y,WU Gang. Concept boundary detection for speeding up SVM[C]//Proceedings of International Conference on Machine Learning. New York,USA:ACM Press,2006:681-688.

[12] SHIN H,CHO S. Neighborhood property based pattern selection for support vector machines[J]. Neural Computation,2007,19(3):816-855.

[13] ANGIULLI F,ASTORINO A. Scaling up support vector machines using nearest neighbor condensation[J]. IEEE Transactions on Neural Networks, 2010, 21(2):351-357.

[14] CHEN Jingnian, ZHANG Caiming, XUE Xiaoping, et al. Fast instance selection for speeding up support vector machines[J]. Knowledge-based Systems, 2013, 45(6):1-7.

[15] HETTICH S,BLAKE C L,MERZ C J. UCI repository of machine learning databases[EB/OL]. [2017-03-21]. http//www. ics. uci. edu/ ~ mlearn/MLRepository. html.

[16] ZHANG Honggang, GUO Jun, CHEN Guang, et al. HCL2000——a large-scale handwritten Chinese character database for handwritten character recognition[C]// Proceedings of the 10th International Conference on Document Analysis and Recognition. Berlin, Germany: Springer,2009:286-289.

编辑 顾逸斐

(上接第 18 页)

[5] 文 琪,彭 宏. 小波变换的离群时序数据挖掘分析[J]. 电子科技大学学报,2005,34(4):556-558.

[6] 侯澍旻. 时序数据挖掘及其在故障诊断中的应用研究[D]. 武汉:武汉科技大学,2006.

[7] 侯澍旻,李友荣,刘光临. 一种基于 KS 检验的时间序列非线性检验方法[J]. 电子与信息学报,2007,29(4):808-810.

[8] WANG Yao, WU Chunguo, JI Zhaohua, et al. Non-parametric change-point method for differential gene expression detection[EB/OL]. [2017-04-14]. https://www. ncbi. nlm. nih. gov/pmc/articles/PMC3104986/.

[9] 王小宜,卢正鼎,凌贺飞. 一个基于小波的时序数据异常探测的新算法[J]. 计算机工程与科学,2005,27(6):83-85.

[10] LIN H D. Automated visual inspection of ripple defects using wavelet characteristic based multivariate statistical approach[J]. Image and Vision Computing, 2007, 25(1):1785-1801.

[11] 刘丹红,张世英. 基于小波神经网络的非线性误差校正模型及其预测[J]. 控制与决策,2006,21(10):1114-1118.

[12] LIN J, KEOGH E, LONARDI S, et al. A symbolic representation of time series, with implications for streaming algorithms[C]//Proceedings of ACM Sigmod Workshop on Research Issues in Data Mining & Knowledge Discovery. New York, USA: ACM Press, 2003:2-13.

[13] 钟清流,蔡自兴. 基于统计特征的时序数据符号化算法[J]. 计算机学报,2008,31(10):1857-1864.

[14] SHARIFZADEH M, AZMOODEH F, SHAHABI C. Change detection in time series data using wavelet footprints[C]//Proceedings of International Symposium on Spatial and Temporal Databases. Berlin, Germany: Springer,2005:127-144.

[15] BRODSKY B E, DARKHOVSKY B S. Nonparametric methods in change point problem[M]. Berlin, Germany: Springer,1993.

[16] ALARCON-AQUINO V, BARRIA J A. Anomaly detection in communication networks using wavelets[J]. IEEE Proceedings Communications, 2002, 148(6):355-362.

编辑 顾逸斐