基于扩展符号聚集近似的水文时间序列异常挖掘

刘　千，朱跃龙，张鹏程
（河海大学 计算机与信息学院，南京 210098）

摘　要：水文时间序列异常挖掘目前大多采用基于距离的方法。为了克服该方法耗时长、计算量大的缺点，采用一种符号化算法，用扩展符号聚集近似对序列符号化表示，再对字符串进行距离度量，并以太湖流域小梅口站逐日水位数据为例进行验证。实验表明该方法的挖掘结果更全面，运算效率很高，更适合处理大规模数据集。

关键词：水文时间序列；异常挖掘；符号化；距离度量
中图分类号：TP399　　**文献标志码**：A　　**文章编号**：1001-3695(2012)12-4479-03
doi:10.3969/j.issn.1001-3695.2012.12.019

Extended symbolic aggregate approximation based anomaly mining of hydrological time series

LIU Qian, ZHU Yue-long, ZHANG Peng-cheng
(*College of Computer & Information*, *Hohai University*, *Nanjing* 210098, *China*)

Abstract: Most of anomaly mining of hydrological time series uses distance based method. Since the method was time-consuming and has a great deal of computation, this paper applied extended symbolic aggregate approximation and then measured distance of strings. It verified the validity of the method by the water level data obtained from Xiaomeikou gauge station in the Taihu Lake. The experimental results show that the method has high efficiency and is more suitable for processing large-scale data sets.

Key words: hydrological time series; anomaly mining; symbolization; distance measure

0　引言

时间序列是一个由随时间变化的序列值或事件数据组成的集合，反映了属性值在时间顺序上的特征，具有数据量大、维数高、更新速度快等特点，在医疗、气象、水文、经济等领域普遍存在。

水文现象是时变现象，这一变化过程称为水文过程。水文数据是对这一过程的离散记录，按其描述的物理量可分为各种类型的水文时间序列[1]，如降雨量、流量、水位等。水文时间序列数据挖掘主要分为相似性搜索、序列模式挖掘和周期分析三类任务，对于异常数据挖掘的研究尚处于起步阶段。水文时间序列的异常可认为是与水文现象的数据模型或一般规律显著不同或不一致的数据对象，而这些数据往往提供了更有价值的水文信息和知识。

本文面向水文时间序列，使用基于扩展符号聚集近似(extended symbolic aggregate approximation，ESAX)进行异常挖掘，实验表明该方法与基于距离的水文时间序列异常挖掘方法相比，在较小的时间复杂度下有较好的挖掘结果，特别对于处理大规模数据集在时间上有较大优势。

1　相关研究

目前，时间序列异常挖掘的方法主要包括：

a)基于距离的方法。由 Leng 等人[2]提出，通过 key points 表征序列，利用二次回归模型实现对序列的不等长划分，最后以动态时间弯曲距离(dynamic time warping，DTW)为基础，给出子序列的 anomaly scores，选取前 k 个最大的值。类似的方法在水文时间序列异常挖掘中已经得到了应用。林森[3]采用了动态时间弯曲距离来度量子序列间的距离，从而得出异常数据；余跃[4]采用动态模式匹配距离(dynamic pattern matching，DPM)来挖掘异常。

b)生物学方法。由 Dasgupta 等人[5]提出，利用生物学免疫系统中自我和异己两个概念，分别映射到时间序列里正常数据和超出允许偏差范围的异常数据，借鉴负选择机制的思想进行检测。这种方法的缺陷在于正常数据越来越多之后，用于负选择的“异己”就会随之减少，最终导致检测过程失败。

c)基于频率的方法。由 Keogh 等人[6]提出，采用后缀树编码时间序列中所有出现的模式，用马尔可夫模型预测未被观测到的模式期望发生概率，然后根据用户给定的阈值来判断模式是否异常。但该方法需要用户给定一个标准的时间序列作为正常的参考值，这是很困难的。

d)支持向量机(support vector machine，SVM)的方法。由 Ma 等人[7]提出，该方法是用支持向量回归(support vector regression，SVR)技术，对历史时间序列建立回归模型，判断新到来的序列点与模型的匹配程度。但是支持向量回归技术理论较复杂，建立模型的过程也比较复杂，实现上有一定难度。

e)基于符号化聚集近似(symbolic aggregate approximation，SAX)的方法。由 Keogh 等人[8]提出，其认为大部分情况下时

收稿日期：2012-04-25；**修回日期**：2012-05-30

作者简介：刘千(1988-)，女，江苏南京人，硕士研究生，主要研究方向为时间序列数据挖掘(liuqian0128@yeah.net)；朱跃龙(1959-)，男，江苏盐城人，教授，博导，主要研究方向为智能数据管理、软件工程；张鹏程(1981-)，男，江苏盐城人，博士，主要研究方向为软件工程、数据挖掘.

间序列中的不一致事件就是该序列中异常程度最大的模式。因此,将原始序列分段聚集近似(piecewise aggregate approximation,PAA)表示后,转换为 SAX 符号,构造符号化数组和单词检索树,查找不一致事件。该方法要求数据必须近似满足正态分布。这种查找时间序列中的不一致事件和时间序列相似性查找要解决的问题相反,特别适合用来检测时间序列的异常情况。

f)TSA-tree。由 Shahabi 等人[9]提出来改善多层次查询的性能,但查询不够准确和全面,有些异常模式无法发现。

上述方法的比较如表 1 所示。

表 1 时间序列异常挖掘主要方法比较

挖掘方法	典型算法	优势	局限性	实验数据
基于距离	Leng Ming-wei[2]	特征点表征序列,二次回归模型对序列不等长划分	大量计算 DTW,耗时	仿真数据和 ECG
生物学	Dasgupta D[5]	思想新颖,不需要先验知识	如果正常数据多样化易导致检测过程失败	Mackey Glass series
基于频率	Keogh E[6]	结合了后缀树字符串匹配、查找能力,效率高	需要给出一组正常值作参考	仿真数据
支持向量机	Ma Jun-shui[7]	相比于神经网络,计算效率高	SVR 理论复杂,建模过程也较复杂	仿真数据和 SFIC
基于 SAX	Keogh E[8]	符号化后可利用字符串查找度量等处理方法,效率高	数据必须近似满足正态分布	ECG
TSA-tree	Shahabi C[9]	改善多层次查询性能	查询不够准确全面	仿真数据和 SP500

综上所述,基于距离的方法时间复杂度较高,效率不能保证;生物学方法无法处理多样化的正常数据;基于频率的方法无法给出一组标准的参考值;支持向量机则对理论和建模过程要求苛刻;TSA-tree 无法保证查找结果的全面性和正确性。因此,为了克服上述方法的缺陷,借鉴基于 SAX 方法中符号化的思想对时间序列进行符号化表示,再对字符串间的距离进行度量,不仅能够保证挖掘结果准确全面,而且能大大降低时间复杂度。

2 基于扩展符号聚集近似挖掘算法

2.1 扩展符号聚集近似

扩展符号聚集近似(ESAX)由 Lkhagva 等人[10]在 2006 年提出,用于克服符号化聚集近似(SAX)[11]方法易丢失分段区间内极值点等重要信息这一缺点,在经济类时间序列子序列查找挖掘任务中有较好的实验结果。水文时间序列的局部极大值和极小值通常比普通数据更能反映变化过程中的一些特性,包含了更多有价值的信息,恰好可以利用该方法进行保留。由于 ESAX 符号化方法必须使得序列满足标准正态分布,但水文时间序列不满足这样的特征,因此在符号化之前,需对原始序列进行标准化。

扩展符号聚集近似的具体步骤如下:

a)采用零—均值方法标准化,对于原始序列 C,将其标准化为序列 $\hat{C}$,其中 u、v 分别为该序列的平均值和标准差:

$$\hat{c}_i = \frac{c_i - u}{v} \tag{1}$$

b)采用 PAA 表示标准化后的序列,对于长度为 n 的序列 $\hat{C}$,通过下式可将其变换为一个长度为 k 的序列 X:

$$x_i = \frac{k}{n} \sum_{j=\frac{n}{k}(i-1)+1}^{\frac{n}{k}i} \hat{c}_j \tag{2}$$

c)将序列 X 划分到 a 个等概率空间,根据 $x_1, x_2, \cdots, x_k$ 的值,把处于同一概率区间的序列值用同一个符号表示,符号总数为 a,即得到一个长度为 k 的符号串。等概率区间的划分见表 2,$\beta_1, \beta_2, \cdots, \beta_9$ 为分位点。相应数值根据标准正态分布表计算得出。例如 $a=3$ 时,每个空间的概率应为 1/3,查找标准正态分布表,$\Phi(0.43) = 0.6664$,$\Phi(-0.43) = 1 - 0.6664 = 0.3336$,即得出分位点的数值。

表 2 符号总数 $a=3,4,\cdots,10$ 时等概率区间的划分

β_i	a							
	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

d)针对步骤 b)中产生的每一个 PAA 分段,找出每一个分段内部的最大值以及最小值,根据表 2 将其也转换为 SAX 符号 S_{max}、S_{min},并记录下它们的位置信息,即横坐标,分别记为 P_{max}、P_{min}。

e)计算 PAA 得到的各段均值所对应的坐标,令任一分段 $\hat{C}_k$,起点为横坐标 S_k,终点为横坐标 E_k,均值所对应的坐标为

$$P_{mid} = \frac{S_k + E_k}{2} \tag{3}$$

f)令任一分段 $\hat{C}_k$ 可转换为一个三维符号向量$\langle S_1, S_2, S_3 \rangle$表示,该向量可由下式得出:

$$\langle S_1, S_2, S_3 \rangle = \begin{cases} \langle S_{max}, S_{mid}, S_{min} \rangle & \text{if } P_{max} < P_{mid} < P_{min} \\ \langle S_{min}, S_{mid}, S_{max} \rangle & \text{if } P_{min} < P_{mid} < P_{max} \\ \langle S_{min}, S_{max}, S_{mid} \rangle & \text{if } P_{min} < P_{max} < P_{mid} \\ \langle S_{max}, S_{min}, S_{mid} \rangle & \text{if } P_{max} < P_{min} < P_{mid} \\ \langle S_{mid}, S_{max}, S_{min} \rangle & \text{if } P_{mid} < P_{max} < P_{min} \\ \langle S_{mid}, S_{min}, S_{max} \rangle & \text{otherwise} \end{cases} \tag{4}$$

通过以上这几个步骤,一个原始序列 C 就转换为了一个由 ESAX 符号所表示的符号串 S。ESAX 的表示方法保留了 SAX 表示方法的优越性,克服了其平滑原序列中局部极大值和极小值的缺点,更能准确地描述时间序列,而且水文时间序列中局部极大值和极小值通常包含了更多有价值的信息,该方法恰恰有效地保留了这部分信息。本文的水文时间序列异常挖掘正是基于 ESAX 的符号化表示。

2.2 距离度量

对于某一段水文时间序列 C 而言,如果其与所属站点其他年份相同时期内的序列间相似性程度较低,或者说,基于某种距离度量方法,C 与这些同时期内的子序列间距离累计之和较大,则 C 可视为一段异常序列。因此,如何对时间序列之间的距离进行度量是对序列符号化之后的重要问题。

本文采用了文献[11]提出的关于符号之间距离的度量方法,通过一个矩阵来描述对应的各个符号间的距离。具体的计算方法如下,其中 i、j 分别表示行、列数,β_n 可参照表 2。

$$dis[i][j] = \begin{cases} 0 & \text{if } |i-j| \leqslant 1 \\ \beta_{max(i,j)-1} - \beta_{min(i,j)} & \text{otherwise} \end{cases} \tag{5}$$

符号总数 $a=5$ 时,即采用 A、B、C、D、E 共五个符号表示序列,各符号间的距离见表 3。例如计算符号 A 和 D 之间的距

离,dis[1][4] $=\beta_3-\beta_1=0.25-(-0.84)=1.09$。

表 3　符号总数 $a=5$ 时各符号之间的距离

	A	B	C	D	E
A	0	0	0.59	1.09	1.68
B	0	0	0	0.5	1.09
C	0.59	0	0	0	0.59
D	1.09	0.5	0	0	0
E	1.68	1.09	0.59	0	0

本文通过上述方法得出了任意两个符号序列之间的距离,进而计算出任意一个符号化后的水文时间序列与其他序列之间的距离之和,下文称其为累积距离,从而根据该距离值的大小挖掘出异常数据。

3　实验分析

采用扩展符号聚集近似以及字符串距离度量方法对具体的水文时间序列进行异常挖掘。关于水文时间序列异常挖掘的研究并不非常广泛,目前大多采用基于距离的方法。文献[4]中的方法是该类方法的典型代表,且思路较为新颖,因此本文选择其作为比较对象,与其中实验情况及算法时间复杂度等进行对比,然后分析实验结果。

3.1　对比分析

实验数据为太湖流域小梅口站 1956—2005 年共 50 年逐日水位数据,与文献[4]相同。实验过程中,对于扩展符号聚集近似算法 B 步骤中 PAA 过程,保证了每段包含六个原始数据,即$[n/k]=6$;算法 C 步骤中符号总数 $a=5$。

本文与文献[4]实验环境不尽相同,具体配置如表 4 所示。

表 4　本文与文献[4]实验环境比较

对比指标	本文	文献[4]
处理器	Intel® pentium® 4	Intel® pentium® 4
CPU 主频	2.66 GHz	2.81 GHz
内存	784 MB	1 GB
操作系统	Windows XP	Windows XP
编译软件	VC ++ 6.0	VC ++ 6.0

根据子序列长度的不同,本文分别采用了与文献[4]相同的查找长度进行实验,即将 50 年水位数据按照各年份划分为 7 ~ 8 月、6 ~ 8 月、6 ~ 9 月、6 ~ 10 月、5 ~ 10 月和 4 ~ 10 月共六个不同的子序列长度,得出程序执行时间,并与其进行对比。具体数据如表 5 和图 1 所示。

表 5　本文与文献[4]查找相同子序列所用时间比较

序号	提取月份	子序列长度	50 年查找长度	本文/s	文献[4]/s
1	7 ~ 8 月	62	3 100	0.055	0.375
2	6 ~ 8 月	92	4 600	0.084	0.688
3	6 ~ 9 月	122	6 100	0.105	1.125
4	6 ~ 10 月	153	7 650	0.134	1.750
5	5 ~ 10 月	184	9 200	0.156	2.688
6	4 ~ 10 月	214	10 700	0.177	3.828

不难看出,本文中所采用的实验环境与文献[4]中虽有一定差距,但是程序执行时间上却有很大的优势,随着数据量的增长,文献[4]中所采用的算法时间复杂度基本随之线性增长;相对而言,本文中所采用的算法增长量微乎其微,可见该方法在处理大型数据集合上更有优势。时间序列符号化之后,虽然在一定程度上降低了数据的精确度,但将大量数据映射到了有限的符号集合,大大减少了后续处理计算的时间。

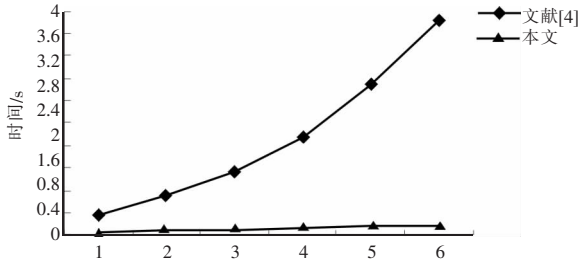

图 1　本文与文献[4]查找相同子序列所用时间比较

3.2　结果分析

实验数据为小梅口站 1956—2005 年共 50 年 7、8 月份逐日水位数据,实验结果如表 6 所示。

表 6　小梅口站 1956—2005 年 7、8 月逐日水位数据异常挖掘结果

排名	年份(7、8 月)	累积距离
1	1991	1044.74
1	1999	1044.74
3	1957	1015.58
4	1987	919.04
5	1958	914.765
5	1978	914.765

在图 2 中,本文选取了与各年份累积距离较小的 2003 年 7、8 月逐日水位数据作为参照,方便与异常挖掘结果作出比较。根据相关资料记载,1991 年和 1999 年太湖流域分别遭遇了百年一遇的特大洪水过程,一般来说太湖流域 7 ~ 8 月汛期水位 3.5 m,但 1991 年和 1999 年均接近 5 m,且整体水位较高;1957 年和 1987 年虽然不及 1991 年和 1999 年涨势凶猛,但大部分时间均超过平均水位,甚至超过 4 m;除洪水外,1958 年和 1978 年还遭遇了两次旱灾过程,在汛期水位较为平缓,大大低于平均水位,且水位甚至不升反降。

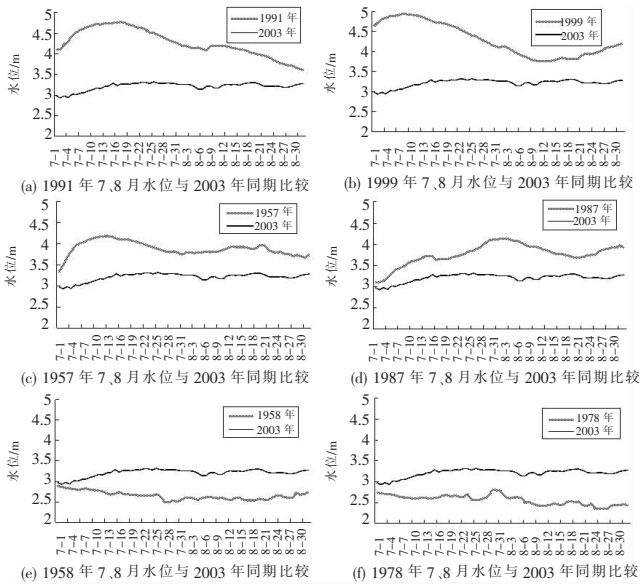

图 2 小梅口站 1956—2005 年 7、8 月逐日水位数据异常挖掘结果

可见,基于扩展符号聚集近似算法能够在较小的时间复杂度下,准确地挖掘出多种不同类型的异常数据,且非常符合实际意义,对于水文现象的观测与研究具有重要意义。

4　结束语

针对水文时间序列异常挖掘,本文运用了基于扩展符号聚集近似算法,以小梅口站 50 年逐日水位数据为例,　(下转第 4502 页)

5 结束语

本文将基因表达谱成功地结合到蛋白质相互作用网络中，构建了高可靠性的加权蛋白质相互作用网络。为了验证加权网络的生物学意义是否比非加权网络更显著，本文用著名的图聚类算法 MCL 从两种网络中预测蛋白质复合物，并对预测出的复合物执行严格的匹配分析。匹配分析的结果显示加权网络有更高的生物显著性。本文的研究验证了“同一物种的不同生物数据之间具有良好的互补性”这一结论。未来将使用数学模型沟通不同数据源，从而结合多数据源对生物信息学的其他研究领域如关键基因预测、致病基因预测和功能模块预测等展开研究。

参考文献：

[1] BADER G D, HOGUE C W. An automated method for finding molecular complexes in large protein interaction networks [J]. BMC Bioinformatics, 2003, 4(2): 59.

[2] HARTWELL L H, HOPFIELD J J, LEIBLER S, *et al*. From molecular to modular cell biology [J]. Nature, 1999, 402(6761): C47-C52.

[3] PEREIR-LEAL J B, ENRIGHT A J, OUZOUNIS C A, *et al*. Detection of functional modules from protein interaction networks [J]. Proteins, 2004, 54(1): 49-57.

[4] RIVES A W, GALITSKI T. Modular organization of cellular networks [J]. Proceedings of National Academy of Sciences, 2003, 100(3): 1128-1133.

[5] SPIRIN V, MIRNY L A. Protein complexes and functional modules in molecular networks [J]. Proceedings of National Academy of Sciences, 2003, 100(21): 12123-12128.

[6] ALTAF-UL-AMIN M, SHINBO Y, MIHARA K, *et al*. Development and implementation of an algorithm for detection of protein complexes in large interaction networks [J]. BMC Bioinformatics, 2006, 7(1): 207.

[7] BLATT M, WISEMAN S, DOMANY E. Superparamagnetic clustering of data [J]. Physical Review Letters, 1996, 76(18): 3251-3254.

[8] BRUN C, CHEVENET F, MARTIN D, *et al*. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network [J]. Genome Biology, 2004, 5(1): 6.

[9] CHEN Jing-chun, YUAN Bo. Detecting functional modules in the yeast protein-protein interaction network [J]. Bioinformatics, 2006, 22(18): 2283-2290.

[10] COLAK R, HORMOZDIAR I F, MOSER F, *et al*. Dense graphlet statistics of protein interaction and random networks [C]//Proc of Pacific Symposium on Biocomputing. 2009: 178-189.

[11] ENRIGHT A J, Van DONGEN S, OUIOUNIS C A, *et al*. An efficient algorithm for large-scale detection of protein families [J]. Nucleic Acids Research, 2002, 30(7): 1575-1584.

[12] GEORGII E, DIETMANN S, UNO T, *et al*. Enumeration of condition-dependent dense modules in protein interaction networks [J]. Bioinformatics, 2009, 25(7): 933-940.

[13] KING A D, PRZULJ N, JURISICA I, *et al*. Protein complex prediction via cost-based clustering [J]. Bioinformatics, 2004, 20(17): 3013-3020.

[14] LOEWENSTEIN Y, PORTUGALY E, FROMER M, *et al*. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space [J]. Bioinformatics, 2008, 24(13): 41-49.

[15] NAVLAKHA S, SCHATI M C, KINGSFORD C, *et al*. Revealing biological modules via graph summarization [J]. Journal Computational Biology, 2009, 16(2): 253-264.

[16] PALLA G, DERENYI I, FARKAS I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818.

[17] SAMANTA M, LIANG S. Predicting protein functions from redundancies in large-scale protein interaction networks [J]. Proceedings of National Academy of Sciences, 2003, 100(22): 12579-12583.

[18] SHARAN R, SUTHRAM S, KELLEY R M, *et al*. Conserved patterns of protein interaction in multiple species [J]. Proceedings of National Academy of Sciences, 2005, 102(6): 1974-1979.

[19] Database of interacting proteins [DB/OL]. http://dip.doe-mbi.ucla.edu/dip/Download.cgi? SM=7/.

[20] TU B P, KUDLICKI A, ROWICKA M, *et al*. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes [J]. Science, 2005, 310(5751): 1152-1158.

[21] PU S, WONG J, TURNER B, *et al*. Up-to-date catalogues of yeast protein complexes [J]. Nucleic Acids Research, 2009, 37(3): 825-831.

(上接第 4481 页)验证了该方法的可行性和有效性，并与基于距离的异常挖掘方法作出对比，突出了该方法保证结果正确全面的前提下，时间复杂度大大降低的优势。目前，符号化方法在水文时间序列异常挖掘领域的研究还不多，今后的研究方向包括对算法的改进，以及该类方法在多维水文时间序列异常挖掘中的运用。

参考文献：

[1] 艾萍，倪伟新. 我国水文数据挖掘技术研究的回顾与展望 [J]. 计算机工程与应用，2003，39(28)：13-17.

[2] LENG Ming-wei, CHEN Xiao-yun, LI Long-jie. Variable length methods for detecting anomaly patterns in time-series [C]//Proc of International Symposium on Computational Intelligence and Design. 2008.

[3] 林森. 时间序列异常检测的研究与应用 [D]. 南京：河海大学，2008.

[4] 余跃. 水文时间序列异常模式挖掘方法的研究与应用 [D]. 南京：河海大学，2011.

[5] DASGUPTA D, FORREST S. Novelty detection in time series data using ideas from immunology [C]//Proc of the 5th International Conference on Intelligent Systems. 1996: 82-87.

[6] KEOGH E, LONARDI S, CHIU B. Finding surprising patterns in a time series database in linear time and space [C]//Proc of SIGKDD'02. 2002: 23-26.

[7] MA Jun-shui, PERKINS S. Time-series novelty detection using one-class support vector machines [C]//Proc of International Joint Conference on Neural Networks. 2003: 1741-1745.

[8] KEOGH E, LIN J, FU A. HOT SAX: finding the most unusual time series subsequence: algorithms and applications [C]//Proc of the 5th IEEE International Conference on Data Mining. 2005: 226-233.

[9] SHAHABI C, TIAN Xiao-ming, ZHAO Wu-gang. TSA-tree: a wavelet-based approach to improve the efficiency of multi-level surprise and trend queries [C]//Proc of the 12th International Conference on Scientific and Statistical Database Management. 2000.

[10] LKHAGVA B, SUZUKI Y, KAWAGOE K. New time series data representation ESAX for financial applications [C]//Proc of the 22nd International Conference on Data Engineering Workshops. 2006: 17-22.

[11] LIN J, KEOGH E, LONARDI S, *et al*. A symbolic representation of time series, with implications for streaming algorithms [C]//Proc of DMKD'03. 2003: 2-11.

[12] HUYNH T, DUONG T. Time series discord discovery based on iSAX symbolic representation [C]//Proc of the 3rd International Conference on Knowledge and Systems Engineering. 2011: 11-18.