文章编号: 1001-0920(2008)10-1109-04

时序数据的动态有界符号化方法

钟清流[1,2], 蔡自兴[1], 陈明权[2], 杨先芬[2]
(1. 中南大学 信息科学与工程学院, 长沙 410083; 2. 湖南大学 计算机与通信学院, 长沙 410082)

摘　要: 提出了一种时序符号化方法. 根据数据集极值来确定最佳字符集及时序数据的划分基准, 通过估算最大压缩比来指导降维, 从而实现了与 SAX 同样的符号化时序转换和相同的距离计算方式. 与 SAX 不同的是, 该时序符号化方法可以有效防止极值信息的丢失, 因而在一些与极值相关的时序分析中有出色的表现.

关键词: 时序分析; 符号化表示; 符号集合近似
中图分类号: TP18　　**文献标识码**: A

Dynamic and limited symbol method for time series data

ZHONG Qing-liu[1,2], *CAI Zi-xing*[1], *CHEN Ming-quan*[2], *YANG Xian-fen*[2]
(1. School of Information Science and Engineering, Central South University, Changsha 410083, China; 2. School of Computer and Communication, Hu'nan University, Changsha 410082, China. Correspondent: ZHONG Qing_liu, E-mail: HNU_ZQL @163.com)

Abstract: A method by which the time series data could be transformed into symbol sequence is presented. The best of all symbols and dividing benchmark are confirmed by means of the extremum about the time series data. Dimension is reduced by estimating the largest compress rate. Thereby, the transformation and distance calculation the same as SAX representation are realized. Different from SAX, it can prevent the information near extremum in a time datasets to be lost, thus has more standout behaves in some time series analysis which depends on extremum.

Key words: Time series analysis; Symbolic method; Symbolic aggregate approximation(SAX)

1 引　　言

符号时间序列分析方法是近年来新兴的一种数据处理方法, 其实质是把维数较高的大规模连续时间序列数据变换为维数较少互不相同值的离散符号序列. 这是一个"粗粒化"过程, 该过程能够从动力系统中快速有效地捕获大尺度的特征, 提取有用定量信息, 从而抑制动力学噪声和测量噪声的影响, 大幅降低计算代价. 计算简单快捷, 是现代时序分析的重要研究方向之一.

符号时序分析的关键是如何有效提取嵌入到时序数据中的相关信息, 这涉及寻找不同原始时序数据的最佳分割方法[1]. 在目前时序符号化的各种方法中, 较著名的包括基于方差[2]、熵[3]、层次聚类[4]、小波[5]、符号假近邻[6]的方法和符号集合近似(SAX)[7]方法. 其中, SAX 方法是 Eammon 在 2003 年提出的一种新型符号化表示方法[7], 在许多应用中都具有良好的性能[8], 因而受到了广泛关注. 其主要优点是: 比其他符号算法更简便、高效; 在符号化过程中实现了减维降噪, 保证在符号空间计算出的两个符号序列距离满足实际两个时间序列距离的下界要求, 即不会出现漏报[7]. 但它只适合于遵循高斯分布且在有限方差范围内有较高分布密度的时序数据, 因为这种基于等概率间隔的近似方法容易丢失一些极值信息, 尤其是位于其符号划分上下边界区的极值信息, 因此不适合以极值点或临界点为重要分析依据的应用场合, 例如金融时序数据分析[8].

在很多时序分析任务中, 时序极值是重要的关键信息, 往往代表着某一趋势的重要转折点. 而 SAX 方法很难避免这种边界区信息丢失问题, 因而不可避免地限制了它在这些任务中的应用. 为此, 本文提出时序数据的动态有界符号化(DLS)方法. 它根据时序数据的极值确定划分的上下边界, 并根

收稿日期: 2007-07-31; **修回日期**: 2007-11-15.
基金项目: 国家基础研究项目(A1420060159).
作者简介: 钟清流(1954—), 男, 湖南桃江人, 副教授, 博士生, 从事人工智能、模式识别等研究; 蔡自兴(1938—), 男, 福建莆田人, 教授, 博士生导师, 从事人工智能、机器人学等研究.

据最大熵确定最佳描述字符集,进而按照上下界范围及最佳字符集数动态地确定划分间隔,从而有效防止边界区的信息隐性丢失. 仿真实验表明,本方案能够适应于许多非高斯分布非线性时序的符号化问题,并提供对极值数据的更多描述信息. 在一些与极值相关的时序相似搜索、分类、异常检测应用中,能够得到比 SAX 更好的结果.

2 DLS 方法

2.1 概 述

本文需要解决的问题是将一段长度为 n 的任意(可能不满足高斯分布)时序数据 $X = x_1, x_2, \cdots, x_n$ 转化成长度为 N 且用字符集 $A = \{A_1, A_2, \cdots\} = \{a, b, \cdots\}$ 表示的符号数据 $\hat{X} = \hat{x}_1, \hat{x}_2, \cdots, \hat{x}_N$. 其中 $N \ll n$.

在转化过程中应尽可能地减少极值信息丢失. 为此,本文采用如下方案:首先计算待转化时序数据的最大压缩比 w(又称最大分段长度,即每个符号所代表的时序数据点数);然后用分段集成近似(PAA[9])方法计算各时段数据的近似表示,进而确定最佳字符集的分量数 A. 与 SAX 方法不同的是,本文所用的符号化方法不使用等概率原则作为离散化的划分依据,而是根据时序数据的极值(最大值或最小值)和最佳字符集规模来动态决定划分间隔的基准,进而实现整个时序序列的符号化.

2.2 时序数据的降维

为了使符号化时序数据能够适用于一般情况,在降维前需将连续时序数据转化为零均值、标准差为 1 的规范化数据;然后用 PAA[9] 方法将标准化时序数据 $X = x_1, x_2, \cdots, x_n$ 按适当的压缩比 w 降维,生成一个 $N = n/w\ (N \ll n)$ 维空间向量 $\overline{X} = \overline{x}_1, \overline{x}_2, \cdots, \overline{x}_N$. 其中 $\overline{x}_i$ 计算如下:

$$\overline{x}_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j. \tag{1}$$

时序数据符号化的目的是在误差允许范围内通过时序数据的减维来降低计算代价,即通过减维操作将维数较大的连续长时序数据压缩为维数较小的离散短符号序列,同时又保证经压缩后的符号时序保留原始时序的重要特征信息,以便在各种时序分析任务中能够以较小的计算代价达到与原始时序分析尽可能相同的效果. 时序数据符号化的关键是寻找最佳压缩比.

2.3 最大压缩比的确定

压缩比相当于一个符号所能代表的时序数据点数. 显然,压缩比越高,所代表的时序数据点数越多,当然对所代表的那段时序描述也就越粗糙(信息丢失越多),计算代价越小. 然而,压缩比越低,所代表的时序数据点数越少,对所代表的那段时序描述也就越精细(信息丢失越少),同时也带来了计算代价变大的问题. 在极端情况下,压缩比为 1,表示符号数据与原始时序等维,即没有任何压缩,当然也不会有任何信息损失,但这违背了时序符号化的初衷. 另一个极端是整个时序串压缩为一个符号表示,压缩比为原始时序的长度. 此时信息几乎全部损失,同样没有意义. 对时序数据符号化有意义的压缩比是指在时序分析误差增加不超过允许值前提下的压缩比. 本文根据大量的仿真实验,发现压缩比与时序数据的变化频率呈相反变化关系,同时也与允许误差相关. 确切的关系目前难以描述,在此仅提出一个经验公式如下:

$$W = \sqrt{2k\theta / F}, \tag{2}$$

$$F = \sum_{i}^{n-1} | x_{i+1,j} - x_{i,j} | / (n-1). \tag{3}$$

其中:W 为最大压缩比,F 为时序数据变化频率,θ 为允许误差,k 为标准常数,n 为数据点总数.

2.4 时序数据的符号化

用 PAA 方法按选定的压缩比 w 降维后,即可根据字符集和时序数据的极值得到描述各个字符所代表数据区间的划分点,进而将已经降维的 PAA 数据离散化成符号数据. 通常,若已经确定了最佳字符集规模为 k,则一个由 k 个字符组成的字符集需要有 $k-1$ 个划分点,因此划分点集可表为

$$C = \{c_1, c_2, \cdots, c_{k-1}\}, \tag{4}$$

其中 C_0 与 C_k 分别为最小值和最大值(下界和上界). 从第 C_{i-1} 个划分点到第 C_i 个划分点的间隔按下式确定:

$$c_i - c_{i-1} = (c_k - c_0)/(k-1). \tag{5}$$

按此方案,可以采用类似于 SAX 的方法[7] 来实现从 PAA[9] 时序到符号化时序的转换.

定义 1 若用 A_i 来表示字符集 A 中的第 i 个字符,则从 PAA 近似 $\overline{X}$ 到符号近似 $\hat{X}$ 的映射可由下式确定:

$$\hat{x}_i = A_j,\ c_{j-1} \leq \overline{x}_i \leq c_j. \tag{6}$$

将所有小于 c_1 的 PAA 时序数据映射为符号 $A = A_{\min}$. 同理,也可将所有大于 c_1 而小于 c_2 区间的 PAA 时序数据映射为符号 A_2,而将所有大于 c_{n-1} 区间的 PAA 时序数据映射为符号 $A_n = A_{\max}$. 若采用英语字符来表示,则可将一个 PAA 时序序列转化为类似于 $\hat{X} = \hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n = a, b, c, \cdots$ 的字符串,从而完成时序数据符号化的全部过程.

3 DLS 符号时序的距离计算

从数据空间向符号空间的映射操作实际上存

在着明显的对应关系,因而可以根据这些对应关系实现符号空间的距离计算及数据空间的欧式距离计算. 回顾符号化操作过程,第 1 步是采用 PAA 减维,将原始数据转化为 PAA 平均值数据,两个原始时序 A 和 B 的欧氏距离可表为

$$D(A,B) = \sqrt{\sum_{i}^{n}(a_i - b_i)^2}, \quad (7)$$

相应的 PAA 时序 $\overline{A}$ 和 $\overline{B}$ 的距离则为

$$DR(\overline{A},\overline{B}) = \sqrt{\frac{n}{N}}\sqrt{\sum_{i}^{N}(\overline{a}_i - \overline{b}_i)^2}. \quad (8)$$

可以证明,两个 PAA 时序 $\overline{A}$ 和 $\overline{B}$ 的距离与原始时序的欧式距离之间存在下面关系[10]:

$$D(A,B) \geq DR(\overline{A},\overline{B}). \quad (9)$$

当进一步将 PAA 数据转化成符号数据后,两个符号时序 A 和 B 间的距离可由式(4) ～ (6) 得到,并由下式计算:

$$\mathrm{dist}(A_i,B_j) = \sqrt{(c_i - c_j)^2} = |c_i - c_j|. \quad (10)$$

表 1 为某一特征时序数据在符号化过程中生成的距离矩阵查找表. $\mathrm{dist}(A_i,B_j)$ 可由表 1 得到,有

$$\mathrm{dist}(A_i,B_j) = \begin{cases} 0, & |A_i - B_j| \leq 1; \\ |c_i - c_{j-1}|, & i,j = 1,2,\cdots,s, \text{ otherwise.} \end{cases} \quad (11)$$

其中 c_i 和 c_j 为与该字符对应的划分点值. 例如,当字符分别为 $A_i = a, B_j = d$ 时,由表 1 可查出 $\mathrm{dist}(a,d) = \mathrm{dist}(d,a) = 1.9455$. 由此可进一步计算它所代表的相应原始时序的最小距离

$$\min \mathrm{dist}(A,B) \geq \sqrt{\frac{n}{N}}\sqrt{\sum_{i}^{N}(\mathrm{dist}(A_i,B_j))^2}. \quad (12)$$

同样可以证明 PAA 时序与符号时序距离间存在下列关系[7]:

$$DR(\overline{A},\overline{B}) \geq \min \mathrm{dist}(A,B). \quad (13)$$

这与 SAX 方法的距离矩阵显然是等价的,因此,本算法实现了与 SAX 相同的维度压缩和下界距离[10,11],能够使用与 SAX 相同的距离计算方法来进行符号时序分析比较.

表 1　某距离矩阵构成的查找表

	a	b	c	d	e
a	0	0	0.97	1.95	2.92
b	0	0	0	0.97	1.95
c	0.97	0	0	0	0.97
d	1.95	0.97	0	0	0
e	2.92	1.95	0.97	0	0

4　仿真实验

为了检验本文算法的有效性,采用表 2 所示的 UCI和 EGC共15个时序数据集,分别用本文提出的 DLS 方法与 SAX方法,将各时序数据集转化成相应的符号时序数据集,并作对比实验. 限于版面,仅给出 15 个数据集中的 6 个. 方法是:分别将被查询序列子串 T_i 与待查询时序数据集 T 转化成相应的符号时序,再用 Brute-Force[12] 算法进行相应的时序分析实验,对比实验的结果如图 1 ～ 图 4 所示.

表 2　用于仿真实验的时序数据集示例

数 据 名	数据集长度	数据源
chfdb-chf01-275	3751	ECG
ECG106-test	1216	ECG
xmitdb-x108	5400	ECG
synthetic-control-data1	600	UCI
synthetic-data1-10k-1	10000	UCI
Coffee-TRAIN	287	UCR
JENKINS8	100	TSDL

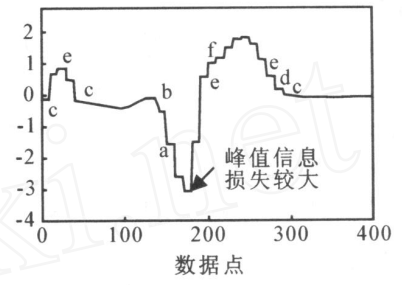

图 1　SAX 符号序列

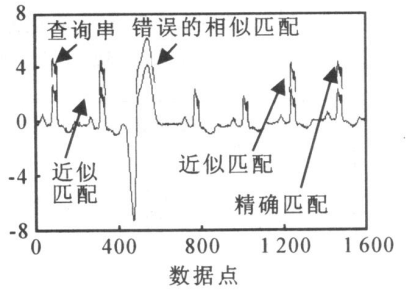

图 2　SAX 符号用于相似查询

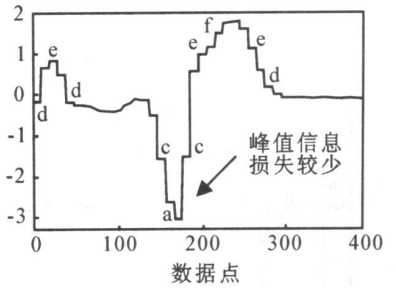

图 3　DLS 符号序列

图 1 表明,尽管位于上下边界区的数据都分别划分为相应的最大、最小字符,但由于这种划分过分粗糙,实际上造成隐性的信息丢失,形成上下边界的信息丢失区,将导致一些时序分析任务中的较大失误. 图 2 显示的是将这一 SAX 符号化结果用于相似查询分析的情况,如果查询串恰好位于边界区,则无

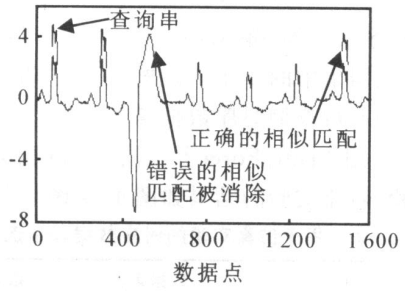

图 4 DLS 符号用于相似查询

论其幅度多大都会被转化为最大或最小字符,尽管图 2 中第 3 个波峰幅度比第 1 个(查询串) 高了几乎一倍,但由于同属于边界区字符,它们都会被转化为同样的字符(在计算时被当作相同的波幅对待),在相似查询中出现错误匹配也在所难免.

4.1 异常检测

当压缩比为 $w = 10$,字符集都取相同规模时,选择一段包含极值字符的正常时序子串 T_i 作为标准串,用经改进的 Brute_Force[12] 算法在整个时序数据集 T 中检测异常. 其分析结果表明,对于大多数据集而言,DLS 检测准确度均优于 SAX. 图 3 和图 4 为采用 ECG 数据源的 chfdb_chf01_275 时序数据集在异常检测时的一个对比实例. 图 4 说明,在一般情况下,DLS 不差于 SAX 符号时序,有些时候甚至更准确. 这是因为异常检测对时序极值信息更敏感,而本文算法恰好加强了极值信息描述的结果.

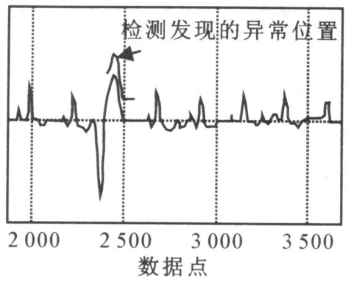

图 5 SAX 符号用于异常检测

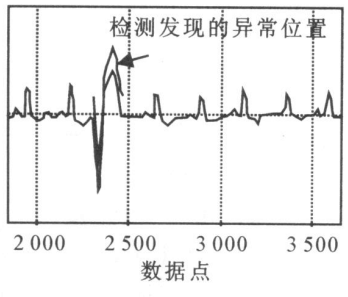

图 6 DLS 符号用于异常检测

4.2 相似查询

通过对比相似查询实验,可以从另一个角度来检验 DLS 符号时序与 SAX 各自的优点. 用 Brute-Force 算法,在时序数据集 T 中(在一定误差范围内) 查找能够与给定长度标准子串 T_i 相匹配的子序列 T_j. 相似查询实验结果表明将会出现两种典型结果:

1) 若标准子串 T_i 包含极值的字符串,则在实现符号化过程采用字符集规模相同的情况下,用 SAX 方法将导致相似查询中的误配率较高,而用 DLS 符号时序误配率较低. 将图 3 与图 1 对比可以发现,由于 DLS 符号化方法防止了峰值信息过多丢失,在边界区(极值区) 的表现好于 SAX,这在不同数据集的类似实验中都得到了证实.

2) 若标准子串 T_i 不包含极值字符,在字符集相同(例如都用 8 个字符) 的情况下,SAX 方法的匹配精度一般高于动态符号时序方法,而动态符号时序要达到与 SAX 方法同样的匹配精度则需要更多的字符,这意味着将要付出更多的计算代价. 表 3 是 15 个不同数据集在相似查询中 20 次实验的平均匹配结果.

表 3 相似查询平均匹配数

符号化方法	相似查询平均匹配数	
	极值查询	非极值查询
SAX	5.81	3.21
DLS	2.53	4.02

在表 3 中极值查询时,SAX 符号时序的平均匹配率远高于 DLS 符号时序,这说明此时它的匹配精度差于后者,因为这中间包含了如图 1 所示的完全错误匹配和一些近似的匹配. 而在非极值查询时,情况相反,此时 DLS 包含的近似匹配多于 SAX,即其匹配精度差于 SAX. 因而两种符号化方法各有不同的应用范围. 从已知的实验来看,DLS 似乎更适合于涉及极值(或转折点) 的相关时序分析任务.

5 结 论

针对 SAX 符号时序方法在边界区信息丢失较多所带来的缺陷,本文提出了 DLS 符号时序方法. 它根据数据集的极值来动态确定最佳字符集和时序数据的划分间隔,通过估算最大压缩比来指导最好的降维比例. 采用 PAA 作降维预处理,从而实现了与 SAX 同样的符号化时序转换和相同的距离计算方式. 但与 SAX 不同的是,DLS 符号时序可以防止位于划分边界区的峰值信息的丢失,因而在一些对峰值信息敏感的时序分析中有比 SAX 更出色的表现. 当然,为此付出的代价是,在相同字符集情况下,它在非边界区内的信息描述不如 SAX 那么精细,为达到同样精度则需增加其字符集,这将增加计算代价. 下一步的工作是要探索既能减少丢失边界区峰值信息,又不影响非边界区描述精度的可能方案.

(下转第 1116 页)

险态度的影响.

5 结 论

本文针对偏好信息以三角模糊数给出的互补判断矩阵,从构造思维的角度考虑了判断矩阵的一致性定义.进而根据三角模糊数互补判断矩阵的完全一致性概念,建立了基于最小方差的非线性规划模型.通过求解该模型讨论了三角模糊数互补判断矩阵的排序问题.算例分析表明了该排序方法是可行而有效的.

参考文献(References)

[1] Satty T L. The analytic herarchy process[M]. New York: Mc Graw-Hill, 1980.

[2] Xu Z S, Wei C P. A consistency improving method in the analytic hierarchy process [J]. European J of Operational Research, 1999, 116(2): 443-449.

[3] Orlovsky S A. Decision-making with a fuzzy preference relation[J]. Fuzzy Sets and Systems, 1978, 1(3): 155-167.

[4] Kacprzyk J. Grouup decision making with a fuzzy linguistic majority[J]. Fuzzy Sets and Systems, 1986, 18(2): 105-118.

[5] Chiclana F, Herrera F, Herrera Viedma E, et al. A classification method of alternatives for multiple preference ordering criteria based on fuzzy majority[J]. J of Fuzzy Mathematics, 1996, 4(4): 128-143.

[6] Van Laarhoven P J M, Pedrycz W. A fuzzy extension of satty's priority theory[J]. Fuzzy Sets and Systems, 1983, 11(1): 229-241.

[7] 姜艳萍, 樊治平. 三角模糊数互补判断矩阵排序的一种实用方法[J]. 系统工程, 2002, 20(2): 89-92.
(Jiang Y P, Fan Z P. A practical ranking method for reciprocal judgment matrix with triangular fuzzy numbers[J]. Systems Engineering, 2002, 20(2): 89-92.)

[8] 巩在武, 刘思峰. 三角模糊数互补判断矩阵的一致性及其排序研究[J]. 控制与决策, 2006, 21(8): 903-907.
(Gong Z W, Liu S F. Consistency and priority of triangular fuzzy number complementary judgment matrix [J]. Control and Decision, 2006, 21(8): 903-907.)

[9] Buckley J J. Fuzzy hierarchy analysis[J]. Fuzzy Sets and Systems, 1985, 17(3): 233-247.

[10] Chang D Y. Applications of the extent analysis method on fuzzy AHP [J]. European J of Operational Research, 1996, 95(3): 649-655.

[11] 徐泽水. 三角模糊数互补判断矩阵排序研究[J]. 系统工程学报, 2004, 19(1): 85-88.
(Xu Z S. On priority method of triangular fuzzy number complementary judgment matrix [J]. J of Systems Engineering, 2004, 19(1): 85-88.)

[12] Liou T S, Wang M J J. Ranking fuzzy numbers with integral value[J]. Fuzzy Sets and Systems, 1992, 50(3): 247-255.

(上接第 1112 页)

参考文献(References)

[1] Daw C S, Finney C E A, Tracy E R. A review of symbolic analysis of experimental data[J]. Review of Scientific Instruments, 2003, 74 (2): 915-930.

[2] Veenman C J, Reinders M J T, Bolt E M, et al. A maximum variance cluster algorithm[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24 (9): 1273-1280.

[3] Chau T, Wong A K C. Pattern discovery by residual analysis and recursive partitioning[J]. IEEE Trans on Knowledge Data Engineering, 1999, 11(6): 833-852.

[4] Kakizawa Y, Shumway R H, Taniguchi N. Discrimination and clustering for multivariate time series [J]. J of American Statistical Assoc, 1999, 93(441): 328-340.

[5] Rajagopalan V, Ray A. Wavelet-based space partitioning for symbolic time series analysis[C]. Proc of IEEE Conf on CDC and ECC. Seville, 2005: 5245-5250.

[6] Kennel M B, Buhl M. Estimating good discrete partitions form observed data: Symbolic false nearest neighbors[J]. Physical Review Letters, 2003, 91(8): 84-102.

[7] Lin J, Keogh E, Lonardi S, et al. A symbolic representation of time series, with implications for streaming algorithms [C]. Proc of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, 2003.

[8] Lkhagva B, Suzuki Y, Kawagoe K. Extended SAX: Extension of symbolic aggregate approximation for financial time series data representation[C]. DEWS2006 Int Sessions Program. 4A-i8.

[9] Keogh E, Chakrabarti K, Pazzani M. Locally adaptive dimensionality reduction for indexing large time series databases [J]. Proc of ACM SIGMOD Conf on Management of Data, 2001, 5(21-24): 151-162.

[10] Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality reduction for fast similarity search in large time series databases[J]. J of Knowledge and Information Systems, 2001, 3: 263-286.

[11] Jessica Lin, Eamonn Keogh, Li Wei, et al. Experiencing SAX: A novel symbolic representation of time series[J]. DMKD J, 2007, 15(2): 107-144.

[12] Keogh E, Lin J, Fu A. Hot SAX: Efficiently finding the most unusual time series subsequence[C]. Proc of the 5th IEEE Int Conf on Data Mining. 2005: 226-233.