

基于 SAX 的时间序列相似性度量方法^{*}

李桂玲^{1a,2}, 王元珍², 杨林权^{1b}, 吴湘宁^{1a}

(1. 中国地质大学 a. 计算机学院; b. 信息工程学院, 武汉 430074; 2. 华中科技大学 计算机科学与技术学院, 武汉 430074)

摘要: 符号化表示是一种有效的时间序列降维技术,其相似性度量是诸多挖掘任务的基础。基于 SAX (symbolic aggregate approximation) 的距离 MINDIST_PAA_iSAX 不满足对称性,在时间序列挖掘中具有局限性,提出了对称的度量 Sym_PAA_SAX,且下界于欧拉距离。在真实数据集和合成数据集上的实验说明下界紧密性较好,相似搜索错报率较低。

关键词: 时间序列; 降维; 相似性度量; 下界

中图分类号: TP311

文献标志码: A

文章编号: 1001-3695(2012)03-0893-04

doi:10.3969/j.issn.1001-3695.2012.03.025

Research on similarity measure for time series based on SAX

LI Gui-ling^{1a,2}, WANG Yuan-zhen², YANG Lin-quan^{1b}, WU Xiang-ning^{1a}

(1. a. School of Computer Science, b. School of Information Engineering, China University of Geosciences, Wuhan 430074, China; 2. School of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074, China)

Abstract: Symbolic approximation is an effective dimensionality reduction technique for time series, its similarity measure is a basis for various mining tasks. MINDIST_PAA_iSAX is a distance function based on symbolic aggregate approximation (SAX), but it does not satisfy symmetry, so it has limitation in mining time series. This paper put forward and proved a symmetric distance measure Sym_PAA_SAX to be lower bounding to Euclidean distance. Experiments on real and synthetic data sets show its better tightness of lower bounding and lower false positives rate in similarity search.

Key words: time series; dimensionality reduction; similarity measure; lower bounding

0 引言

时间序列是指随着时间的先后顺序而变化的一系列数据,是一类多维的复杂类型数据,目前广泛地存在于金融、科学、工程、医疗等领域。例如某股票某段时间内的开盘价和收盘价的走势、就医者的心电图数据、网络监控中的网络流量、自然现象观测中的大气、温度、风、地震等数据,均是时间序列。

近年来,时间序列数据的挖掘吸引了越来越多研究者的关注,相似性度量是其中的一个重要子问题。所谓相似性度量是指如何衡量时间序列之间的相似性和相似程度,合理的相似性度量是相似搜索、聚类、分类、异常检测、主题发现等诸多挖掘任务的基础。

对于原始时间序列,经典的相似性度量有欧拉距离 (Euclidean distance, ED)^[1] 和动态时间弯曲 (dynamic time warping, DTW)^[2] 两种。欧拉距离使用广泛,优点是公式简单、易于快速计算,但对噪声很敏感,不能处理不等长序列,不能捕捉具有伸缩性或弯曲的相似模式。动态时间弯曲可处理不等长序列,允许序列的偏移和扭曲,但计算的时间代价较大。

由于时间序列数据具有海量、高维的特性,研究者对时间

序列作降维处理,进行近似表示。代表性的时间序列的近似表示有分段聚集近似 (piecewise aggregate approximation, PAA)^[3]、分段线性近似 (piecewise linear approximation, PLA)^[4]、符号聚集近似 (SAX)^[5]、可索引符号聚集近似 (indexable SAX, iSAX)^[6]、扩展的符号聚集近似 (extended SAX, ESAX)^[7]、分段线性聚集近似 (piecewise linear aggregate approximation, PLAA)^[8]、无限长时间序列的分段线性拟合 (infinite time series piecewise linear fitting, ITSPFL)^[9] 等。

基于时间序列的近似表示提出了相应的距离公式。Keogh 等人基于 PAA 表示提出 DR^[3] 和 LB_PAA^[10], Lin 等人^[5] 基于 SAX 提出 MINDIST, 兰妥等人^[11] 基于 ESAX 提出 ESAX 统计向量空间法的相似性度量, Huang 等人^[8] 在 PLAA 基础上提出基于子段均值和最佳拟合直线的斜率距离公式, Shieh 等人^[6] 基于 iSAX 提出 MINDIST_PAA_iSAX。

符号化表示是一种有效的离散化的时间序列降维方法。SAX 和 iSAX 均是允许降维和支持下界的简单高效的符号表示法。研究发现, MINDIST_PAA_iSAX 是一种不对称的距离函数,在时间序列的挖掘中具有局限性。本文基于 SAX 表示,提出一种新的距离度量 Sym_PAA_SAX。Sym_PAA_SAX 考虑相

收稿日期: 2011-08-24; **修回日期:** 2011-10-09 **基金项目:** 湖北省自然科学基金资助项目 (2009CDB226); 中央高校基本科研业务费专项资金资助项目 (CUGL100243)

作者简介: 李桂玲 (1979-), 女, 湖北人, 讲师, 博士研究生, 主要研究方向为数据挖掘和知识发现 (lgldec@yahoo.com.cn); 王元珍 (1945-), 女, 教授, 博导, 主要研究方向为数据挖掘、数据库理论与技术; 杨林权 (1979-), 男, 讲师, 博士研究生, 主要研究方向为人工智能、数据挖掘; 吴湘宁 (1972-), 男, 副教授, 主要研究方向为数据仓库和数据挖掘。

似性度量中两个序列的对等地位,具有对称性,满足下界定理。

1 时间序列的符号化表示

1.1 SAX和iSAX表示

Keogh 等人在 PAA 基础上提出时间序列的符号化离散表示 SAX^[5]。已知时间序列 $C = \{c_1, c_2, \dots, c_n\}$, 对其 SAX 表示的具体步骤如下:

a) 规格化。将时间序列 C 规格化为均值为 0、标准差为 1 的标准序列 C' 。

b) PAA 降维。对 C' 进行 PAA 表示, 得到 $\bar{C} = \{\bar{c}_1, \bar{c}_2, \dots, \bar{c}_w\}$ 。其中 w 为时间序列的 PAA 表示的长度。

c) 离散化。选定字母集大小即基大小 a , 根据字母集大小在高斯分布表(如表 1)中查找区间的系列分裂点 β_i , 将 PAA 表示的均值映射为对应的字母, 最终离散化为字符串 $\hat{C} = \{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_w\}$ 。

表 1 基从 2 到 6 的分裂点

β_i	a				
	2	3	4	5	6
β_1	0.00	-0.43	-0.67	-0.84	-0.97
β_2		0.43	0	-0.25	-0.43
β_3			0.67	0.25	0.00
β_4				0.84	0.43
β_5					0.97

SAX 表示具有计算简单、高效的优点, 并且允许降维、支持下界函数, 一经提出, 便成为一种非常受欢迎的时间序列降维表示法。Shieh 等人^[6]在 SAX 基础上提出 iSAX。iSAX 不同于 SAX 主要在于用二进制形式替代字母表示每个分段。例如已知长度 16 的时间序列 $C = \{-1.0, -0.85, -0.75, -0.6, 0.1, 0.5, 0.2, 0.4, -0.1, -0.2, -0.4, -0.1, 0.7, 0.95, 1.05, 0.9\}$, 图 1、2 为时间序列 C 的两种符号化表示示意图, 压缩比为 4, 基为 4。 C 的 SAX 和 iSAX 表示分别为 $C_{SAX} = \{a, c, b, d\}$, $C_{iSAX} = \{3^4, 1^4, 2^4, 0^4\}$ 。

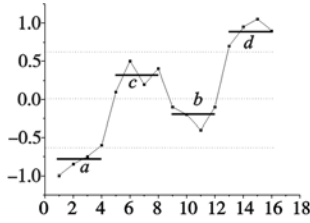


图1 时间序列C的SAX表示

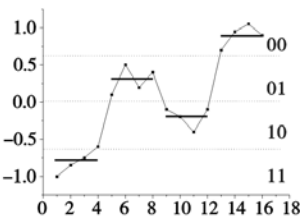


图2 时间序列C的iSAX表示

1.2 MINDIST_PAA_iSAX

Shieh 等人在文献^[6]中定义查询序列的 PAA 表示和待查询空间中 iSAX 表示的下界距离 MINDIST_PAA_iSAX。给定时间序列 T 的 PAA 表示 T_{PAA} , 时间序列 S 的 iSAX 表示 S_{iSAX} , 其中 $|T| = |S| = n$, $|T_{PAA}| = |S_{iSAX}| = w$ 。追溯源自 PAA 值的 S_{iSAX} 的第 j 个基值, 两个分裂点 β_L, β_U 间的 v 值, 其中 $\beta_L < v \leq \beta_U, 1 \leq j \leq w$ 。定义距离如下:

$$\text{MINDIST_PAA_iSAX}(T_{PAA}, S_{iSAX}) = \sqrt{\frac{n}{w}} \sum_{i=1}^w \begin{cases} (\beta_{Li} - T_{PAAi})^2 & \text{if } \beta_{Li} > T_{PAAi} \\ (\beta_{Ui} - T_{PAAi})^2 & \text{if } \beta_{Ui} < T_{PAAi} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

该距离下界于欧拉距离, 然而并不满足对称性, 下面举例证明。假设 $S_{PAA} = \{-0.9, -0.7, 0.3, 0.8\}$, $T_{PAA} = \{-0.8, -0.1, 0.3, 0.8\}$, 选取基为 4, 分裂点依次为 $\{-0.67, 0, 0.67\}$, 可得 $S_{iSAX} = \{3^4, 3^4, 1^4, 0^4\}$, $T_{iSAX} = \{3^4, 2^4, 1^4, 0^4\}$ 。按式(1)计算发现, $\text{MINDIST_PAA_iSAX}(T_{PAA}, S_{iSAX}) = 0.3249$, $\text{MINDIST_PAA_iSAX}(S_{PAA}, T_{iSAX}) = 0.0009$, 两者不等, 该公式不满足对称性。分析其原因在于, 该公式中未置 T 和 S 于同等地位。

2 对称的距离度量 Sym_PAA_SAX

2.1 公式的提出

针对 MINDIST_PAA_SAX 的不对称性, 考虑 S 和 T 在距离计算中的地位对等性, 基于 SAX 表示提出对称的距离度量 Sym_PAA_SAX。该公式定义如下:

$$\text{Sym_PAA_SAX}(S, T) = \sqrt{\frac{n}{w}} \sum_{i=1}^w \begin{cases} (\max(T_{PAAi} - \beta_{SUi}, \beta_{TLi} - S_{PAAi}))^2 & \beta_{TLi} \geq \beta_{SUi} \\ (\max(\beta_{SLi} - T_{PAAi}, S_{PAAi} - \beta_{TUi}))^2 & \beta_{SLi} \geq \beta_{TUi} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

根据 T_{PAAi} 和 S_{PAAi} 在 SAX 表示的分裂点区间中的相对位置划分为以下三种情况:

- a) T_{PAAi} 与 S_{PAAi} 属于同一区间时, 两者相对距离为 0。
- b) T_{PAAi} 与 S_{PAAi} 不在同一区间且 T_{PAAi} 在 S_{PAAi} 上方, 即 $\beta_{TLi} \geq \beta_{SUi}$ 时, 两者相对距离取 $T_{PAAi} - \beta_{SUi}$ 和 $\beta_{TLi} - S_{PAAi}$ 较大者。
- c) T_{PAAi} 与 S_{PAAi} 不在同一区间且 S_{PAAi} 在 T_{PAAi} 上方, 即 $\beta_{SLi} \geq \beta_{TUi}$ 时, 两者相对距离取 $\beta_{SLi} - T_{PAAi}$ 与 $S_{PAAi} - \beta_{TUi}$ 较大者。

设 $(-\infty, -0.67), [-0.67, 0), [0, 0.67), [0.67, +\infty)$ 为 SAX 表示法的分裂区间集合, $S_{PAA} = \{-0.8, 0.3, -0.2, 0.8\}$, $T_{PAA} = \{-0.7, -0.1, 0.5, 0.3\}$, S 和 T 的 SAX 表示如图 3 所示。 S 和 T 各有四个子段, 按以上三种情况分析计算, 第一个子段落在相同区间, 两者相对距离为 0; S 的第二个子段在 T 的上方, 取 $\max(0 - (-0.1), 0.3 - 0) = 0.3$; T 的第三个子段在 S 的上方, 取 $\max(0.5, 0.2) = 0.5$; S 的第四个子段在 T 的上方, 取 $\max(0.67 - 0.3, 0.8 - 0.67) = 0.37$ 。

2.2 对称性证明

从图 3 可以看出, Sym_PAA_SAX 是对称的距离度量。下面证明其对称性。

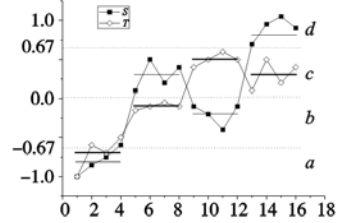


图3 时间序列S、T及其SAX示意图

定理 1 Sym_PAA_SAX 满足对称性。

证明 由式(2)可得

$$\text{Sym_PAA_SAX}(T, S) = \sqrt{\frac{n}{w}} \sum_{i=1}^w \begin{cases} (\max(S_{PAAi} - \beta_{TUi}, \beta_{SLi} - T_{PAAi}))^2 & \beta_{SLi} \geq \beta_{TUi} \\ (\max(\beta_{TLi} - S_{PAAi}, T_{PAAi} - \beta_{SUi}))^2 & \beta_{TLi} \geq \beta_{SUi} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

比较式(2)和(3), 显然

$$\text{Sym_PAA_SAX}(S,T) = \text{Sym_PAA_SAX}(T,S) \tag{4}$$

$$c_i' = (c_i - \mu) / \sigma \tag{10}$$

2.3 下界性证明

对原始时间序列降维处理后,在降维的特征空间中查询时容易出现两类问题,即错报(false positives)和漏报(false dismissals)^[12]。

错报是指原始空间中两点距离大于阈值 ε ,但是降维后的映射空间中两点距离小于阈值 ε ,导致报告了一个错误的查询结果,即

$$\begin{aligned} \exists c_i, c_j \in C, \overline{c_i}, \overline{c_j} \in \overline{C}, \varepsilon > 0 \\ D_{\text{feature}}(\overline{c_i}, \overline{c_j}) < \varepsilon \Rightarrow D_{\text{object}}(c_i, c_j) \geq \varepsilon \end{aligned} \tag{5}$$

对于错报的结果,可回到原始空间中检测,如果发现原始空间中的点距离大于阈值 ε ,则将其从查询结果中剔除。

漏报是指原始空间中两点距离小于阈值 ε ,但是降维后的特征空间中两点距离大于阈值 ε ,导致没有报告该查询结果,即

$$\begin{aligned} \exists c_i, c_j \in C, \overline{c_i}, \overline{c_j} \in \overline{C}, \varepsilon > 0 \\ D_{\text{feature}}(\overline{c_i}, \overline{c_j}) \geq \varepsilon \Rightarrow D_{\text{object}}(c_i, c_j) < \varepsilon \end{aligned} \tag{6}$$

为了保证无漏报,Faloutsos 等人^[12]提出下界(lower bounding)定理,即

$$D_{\text{feature}}(F(O_1), F(O_2)) \leq D_{\text{object}}(O_1, O_2) \tag{7}$$

其中, O_1 和 O_2 是原始空间的两个对象,度量的距离函数是 $D_{\text{object}}()$; $F(O_1)$ 和 $F(O_2)$ 是降维处理的特征提取函数,度量的距离函数是 $D_{\text{feature}}()$ 。

下面证明 Sym_PAA_SAX 下界于真实的欧拉距离。

定理 2 Sym_PAA_SAX 满足欧拉距离的下界定理。

证明

$$\begin{aligned} \text{Sym_PAA_SAX}(S,T) &\leq \\ \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^w (S_{\text{PAA}i} - T_{\text{PAA}i})^2} &\leq \\ \sqrt{\sum_{i=1}^n (S_i - T_i)^2} = ED(S,T) \end{aligned} \tag{8}$$

即有

$$\text{Sym_PAA_SAX}(S,T) \leq ED(S,T) \tag{9}$$

3 实验评价

3.1 实验环境与实验数据说明

使用 C++ 语言进行编程。实验环境为:Windows XP SP2 Professional 操作系统,1 GB 内存,Pentium Dual 1.73 GHz CPU, 120 GB 硬盘,Visual C++ 6.0 编译环境。

实验数据选用两类数据集:random walk 和 Morgan Stanley stock。Random walk 为典型的合成数据集^[13],大小为 65 536。具体为连续的模拟对称随机游走,每次的增量 $y(s+t) - y(t)$ 是高斯分布 $N(0, t^2)$,不相交间隔的增量是独立的。Morgan Stanley stock 来源于雅虎财经的真实数据集^[14],为摩根斯坦利公司股票的历史交易数据,大小为 4 096,具体为摩根斯坦利公司的股票从 1993 年 2 月 23 日(开盘日)到 2009 年 5 月 27 日的交易日收盘价(close price)。开盘日的收盘价为 31,这段交易日内最低收盘价为 9.2,最高收盘价为 142.75。对两个数据集进行子序列的相似搜索。实验的评价标准包括下界紧密性、错报率和计算时间三个方面。

为方便实验处理,对数据集 C 规格化为均值为 0、标准差为 1 的序列 C' 。先求出序列 C 的均值 μ 和标准差 σ ,规格化如式(10)所示。

3.2 下界紧密性比较

下界距离的好坏通常用紧密性(tightness of lower bound, TLB)^[7]来衡量,TLB 定义为

$$\text{TLB} = \frac{\text{lower bound distance}}{\text{true euclidean distance}} \tag{11}$$

由紧密性定义可以看出,TLB 取值范围在 0~1 之间,值越大越好。紧密性越接近 1,说明下界距离越接近于真实的距离度量,就更能缩小误差。

比较本文提出的 Sym_PAA_SAX 与 DR^[3]、LB_PAA^[10]、MINDIST^[5]、MINDIST_PAA_iSAX^[6]的紧密性,实验结果如图 4、5 所示。

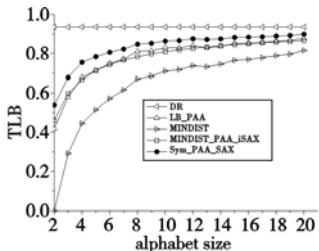


图4 random walk 下界紧密性比较

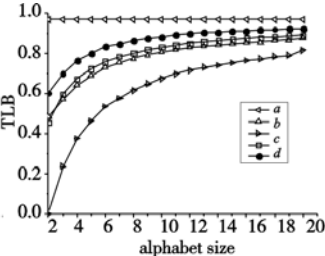


图5 Morgan Stanley stock 下界紧密性比较

图 4、5 分别反映了两类数据集上五种距离的 TLB 随 SAX 表示中基大小变化的趋势。由图 4 和 5 可以看出,对于合成数据集和真实数据集,除 DR 外,其余四种距离均随着字母集基的增大,下界紧密性越好。此外,Sym_PAA_SAX 的紧密性比 LB_PAA、MINDIST、MINDIST_PAA_iSAX 更好,尽管次于 DR,但随着字母集基的增大,逐渐接近于 DR。

3.3 错报率比较

如果下界距离满足非负性和对称性,即自身是距离度量,则可无缝地嵌入到时间序列的相似计算中。显然 Sym_PAA_SAX 满足非负性和对称性。将 Sym_PAA_SAX 用于子序列的相似搜索中,并与 MINDIST 进行比较。由于 Sym_PAA_SAX 满足下界定理,可以保证无漏查。对于错查问题,采用错报率(false positives rate)来评估。错报率定义为

$$\text{false_positives_rate} = \frac{|R| - |R \cap C|}{|R|} \times 100\% \tag{12}$$

其中: R 为算法报告的相似子序列集合, C 为实际正确的相似子序列集合。

对两个数据集进行了多组测试,部分实验结果如表 2、3 所示。表中有三个参数, a 为字母集基, cr 为压缩比(compression ratio),即降维时原始子序列压缩映射为一个 PAA 子段的长度, ε 为阈值。

从表 2、表 3 可以发现,采用 Sym_PAA_SAX 进行距离度量时相似搜索的错报率比 MINDIST 低,原因在于 MINDIST 度量时相同字母及相邻字母的距离定义为 0,降低了准确性,错报

率相对较高,而 Sym_PAA_SAX 能更近似准确地度量序列的相似性,错报率较低。结果显示存在错查,分析原因是降维表示的空间中点距仍较大,未能剔除不合要求的序列。

表 2 random walk 错报率比较

false positives rate/%		parameter setting
MINDIST	Sym_PAA_SAX	
48.06	23.06	$a = 6, cr = 4, \varepsilon = 10$
40.5	19.9	$a = 8, cr = 4, \varepsilon = 10$
34.73	16.18	$a = 12, cr = 4, \varepsilon = 10$
28.64	14.24	$a = 15, cr = 4, \varepsilon = 10$

表 3 Morgan Stanley stock 错报率比较

false positives rate/%		parameter setting
MINDIST	Sym_PAA_SAX	
39.32	27.25	$a = 6, cr = 4, \varepsilon = 27$
39.32	22.12	$a = 8, cr = 4, \varepsilon = 27$
34.43	15.87	$a = 12, cr = 4, \varepsilon = 28$
34.43	13.4	$a = 15, cr = 4, \varepsilon = 28$

3.4 计算时间比较

使用 CPU clock 来评估三种基于 SAX 的距离公式,即 MINDIST、MINDIST_PAA_iSAX、Sym_PAA_SAX,在两类数据集上进行相似搜索所消耗的时间。重复实验多次,结果中的消耗时间是相同参数下多次实验的平均值。

实验结果如图 6、7 所示。图 6 是三种距离在 random walk 数据集上的计算时间比较结果,压缩比为 32,阈值为 10,时间单位为 ms。图 7 是三种距离在 Morgan Stanley stock 数据集上的计算时间比较结果,压缩比为 24,阈值为 27,时间单位为 ms。

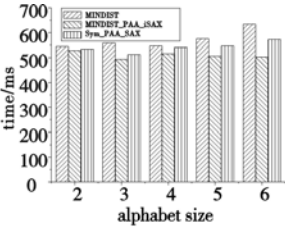


图6 random walk 计算时间比较

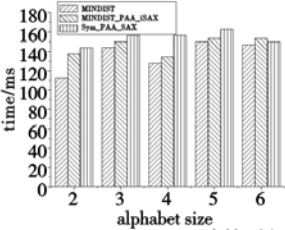


图7 Morgan Stanley stock 计算时间比较

图 6 说明,对于合成数据集 random walk, Sym_PAA_SAX 的计算时间比 MINDIST 要少,稍大于 MINDIST_PAA_iSAX;图 7 说明,对于真实数据集 Morgan Stanley stock, Sym_PAA_SAX 的计算时间稍大于 MINDIST、MINDIST_PAA_iSAX,但比较接近,它们的计算时间的差值是可接受的。

综合上述比较,在两类数据集上, Sym_PAA_SAX 的计算时间在可接受的范围内稍大于 MINDIST_PAA_iSAX,但是 MINDIST_PAA_iSAX 不对称, Sym_PAA_SAX 是对称的度量,且紧密性比 MINDIST_PAA_iSAX 好; Sym_PAA_SAX、MINDIST 同属于对称的度量,计算时间依赖于数据集,但是前者的错报率比后者低,紧密性更好。

4 结束语

本文研究基于 SAX 降维表示的时间序列的相似性度量问题。已有的距离函数 MINDIST_PAA_iSAX 不满足对称性,考虑到两个序列的对等地位,提出了新的相似性度量 Sym_PAA_SAX,并且证明其对称性和下界性。在真实数据集和合成数据集上实验,并与几种典型的距离公式比较,结果表明该度量的下界紧密性好,相似搜索中错报率较低。

实验中同时发现,相似搜索的结果中还存在一定程度上的错报,且计算时间有待优化。下一步研究的方向是如何进一步降低错报率,减少计算时间,并将该度量应用于时间序列的聚类、主题发现、异常检测等挖掘任务中。

参考文献:

[1] CHIU B, KEOGH E, LONARDI S. Probabilistic discovery of time series motifs[C]//Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003:493-498.

[2] BERNDT D J, CLIFFORD J. Using dynamic time warping to find patterns in time series[C]//Proc of KDD Workshop. [S. l.]: AAAI Press, 1994:359-370.

[3] KEOGH E, CHAKRABARTI K, PAZZANI M, et al. Dimensionality reduction for fast similarity search in large time series databases[J]. Knowledge and Information Systems, 2000, 3(3):263-286.

[4] SHATKAY H, ZDONIK S. Approximate queries and representations for large data sequences[C]//Proc of the 12th International Conference on Data Engineering. Washington DC: IEEE Computer Society, 1996:536-545.

[5] 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. New York: ACM Press, 2003:2-11.

[6] SHIEH J, KEOGH E. iSAX: disk-aware mining and indexing of massive time series datasets[J]. Data Mining and Knowledge Discovery, 2009, 19(1):24-57.

[7] 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. Washington DC: IEEE Computer Society, 2006:17-22.

[8] HUANG N Q V, ANH D T. An improvement of PAA for dimensionality reduction in large time series databases[C]//Lecture Notes in Computer Science, vol 5351. Berlin: Springer, 2008:698-707.

[9] 闫秋艳, 夏士雄. 一种无限长时间序列的分段线性拟合算法[J]. 电子学报, 2010, 38(2):443-448.

[10] KEOGH E. Exact indexing of dynamic time warping[C]//Proc of the 28th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 2002:406-417.

[11] 兰安, 江戈, 张东站. 基于 ESAX 表示的时间序列相似性度量[J]. 计算机科学, 2008, 35(10):87-89.

[12] FALOUTSOS C, RANGANATHAN M, MANOLOPOULOS Y. Fast subsequence matching in time-series databases[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1994: 419-429.

[13] UCR time series dataset archives[EB/OL]. http://www.cs.ucr.edu/~eamonn/time_series_data.

[14] Yahoo! Finance[EB/OL]. <http://finance.yahoo.com>.