

DTW 提前终止率的估算方法研究

李正欣¹ 张凤鸣¹ 李俊奎² 张晓丰¹

(空军工程大学工程学院 西安 710038)¹ (华中科技大学数据库与多媒体研究所 武汉 430074)²

摘 要 提前终止在提高时间序列相似性搜索的效率、减少冗余计算等方面起着重要作用。目前缺少一种有效的理论分析方法以相对较小的计算代价估算出提前终止策略对 DTW 计算的优化程度。分析了 DTW 提前终止产生的机理,提出了一种 DTW 提前终止率的估算模型,并对其进行了有效性验证。实验结果表明,该方法能够有效地估算出 DTW 距离的提前终止率,且估算精度优于 EaEst 方法。

关键词 时间序列,动态时间弯曲,提前终止,概率

中图法分类号 TP311 **文献标识码** A

Research on Estimate Method of DTW Early Abandon Ratio

LI Zheng-xin¹ ZHANG Feng-ming¹ LI Jun-kui² ZHANG Xiao-feng¹

(Engineering Institute, Air Force Engineering University, Xi'an 710038, China)¹

(Research Institute of Database & Multimedia, Huazhong University of Science & Technology, Wuhan 430074, China)²

Abstract Early abandon is of great importance in improving efficiency of time series similarity search and reducing the redundant computations. However, previous works are focused on empirical experimental to estimate the effects of early abandon and theoretical analysis method is not available. The mechanism of DTW early abandon was analyzed, and a model of estimating DTW early abandon ratio was proposed, then, experiments were made to testify its validity. The results of experiments show that the proposed method can effectively estimate DTW early abandon ratio, and perform better in precision than that of EaEst method.

Keywords Time series, Dynamic time warping, Early abandon, Probability

1 引言

提前终止(Early Abandon)是在搜索时间序列相似性时采用的一项技术,其递进地计算距离,不断与给定的差异阈值进行比较。一旦发现累积距离超过差异阈值,则断定最终的距离必将超过差异阈值,此时停止计算^[1]。提前终止能够及时停止在不符合条件序列上的相似性计算,从而节省计算资源,提高搜索效率。

动态时间弯曲(Dynamic Time Warping, DTW)是语音识别中广泛使用的技术。Berndt 和 Clifford 在 1993 年将其引入到时间序列的相似性研究中^[2],之后有学者将其扩展到多元时间序列的相似性度量^[3]。DTW 距离定义了时间序列之间的最佳对齐匹配关系,支持时间轴的伸缩和弯曲,已成为时间序列数据挖掘领域中一种重要的相似性匹配方法,但其计算复杂度较高^[4]。为了降低计算复杂度,文献[5]利用提前终止策略,对 DTW 的计算过程进行优化,提出了 Optimized Naive-Scan 方法。文献[6]从 DTW 距离矩阵的性质出发,提出了精确 DTW 下的提前终止方法。文献[7]提出一种名为 EA-DTW 的方法,其用于加速 DTW 距离的计算。在计算累积距

离矩阵中每个方格的距离时都判断其是否超过差异阈值,一旦超过,则提前终止其余相关方格的距离计算。

以往的研究都是利用提前终止策略优化 DTW 计算,而具体的优化程度只能通过 DTW 距离的实际计算得出。目前缺少一种有效的理论分析方法以相对较小的计算代价估算出提前终止策略对 DTW 计算的优化程度,从而减少盲目计算量。文献[1]曾提出一种提前终止率的估算方法(记为 EaEst),然而该方法在理论分析方面存有疏漏,估算精度有待提高。

本文分析了提前终止产生的机理,给出了 DTW 提前终止率的计算方法,提出了一种 DTW 提前终止率的估算模型,并通过实验验证了模型的有效性。

2 DTW 提前终止的机理分析

定义 1(DTW 距离) 时间序列 $Q = \langle q_1, q_2, \dots, q_n \rangle$ 及 $C = \langle c_1, c_2, \dots, c_m \rangle$ 的 DTW 距离定义如式(1)所示, $D_{\text{base}}(q_i, c_j)$ 表示点 q_i 和 c_j 的基距离,通常用 Minkowski 距离表示。DTW 距离实质是确定序列的对齐匹配关系,使得点对距离之和最小,详细内容见文献[2]。

到稿日期:2011-06-13 返修日期:2011-09-18

李正欣(1982—),男,博士生,主要研究方向为信息系统工程与智能决策、数据挖掘, E-mail: lizhengxin_2005@163.com; 张凤鸣(1963—),男,教授,博士生导师,主要研究方向为信息系统工程与智能决策; 李俊奎(1981—),男,博士,主要研究方向为数据挖掘、机器学习; 张晓丰(1978—),男,博士,讲师,主要研究方向为信息系统工程与智能决策。

$$D_{\text{dtw}}(Q, C) = D_{\text{base}}(q_1, c_1) + \min \begin{cases} D_{\text{dtw}}(Q, C[2:-]) \\ D_{\text{dtw}}(Q[2:-], C) \\ D_{\text{dtw}}(Q[2:-], C[2:-]) \end{cases} \quad (1)$$

定义 2(范围查询)^[8] 给定查询序列 Q , 距离度量 D_{dtw} , 距离差异阈值 $\epsilon (\epsilon > 0)$, 若序列 C 满足 $D_{\text{dtw}}(Q, C) \leq \epsilon$, 则称 C 为查询结果。

提前终止是范围查询中的一项技术, 在 $D_{\text{dtw}}(Q, C) \leq \epsilon$ 的计算过程中, 不断地用当前累积距离与差异阈值进行比较, 一旦超过阈值, 则停止计算。下面均是在 DTW 范围查询的条件下进行讨论。

定义 3(有效路径) 若点对距离 $D_{\text{base}}(q_i, c_j)$ 包含在最终的 $D_{\text{dtw}}(Q, C)$ 中, 则 (q_i, c_j) 对最终的距离有贡献, 这样的点对按顺序形成的序列称为有效路径。有效路径上的点对距离记为 $d_1, d_2, \dots, d_n$, 则有式(2)成立。

$$D_{\text{dtw}}(Q, C) = \sum_{i=1}^n d_i, d_i \geq 0 \quad (2)$$

定义 4(有效路径溢出) 在阈值为 ϵ 的范围查询时, 如果 $\sum_{i=1}^k d_i > \epsilon (1 \leq k \leq n)$, 则称有效路径在 k 处溢出。

定理 1(提前终止原理) 在对 Q 进行阈值为 ϵ 的范围查询中, 求解 $D_{\text{dtw}}(Q, C)$ 时, 如果有效路径在 k 处溢出, 则序列 C 不可能是最终的结果序列, 停止计算将不会对查询结果产生影响。

证明: 因为有效路径在 k 处溢出, 所以 $\sum_{i=1}^k d_i > \epsilon (1 \leq k \leq n)$; 由定义 3 得 $D_{\text{dtw}}(Q, C) = \sum_{i=1}^n d_i$, 且有 $\sum_{i=1}^n d_i = \sum_{i=1}^k d_i + \sum_{j=k+1}^n d_j$ 。由于距离的非负性, 有 $\sum_{j=k+1}^n d_j \geq 0$, 因此 $D_{\text{dtw}}(Q, C) > \epsilon$ 。所以, 无论是否对余下的 $n-k$ 个点对距离进行累加, C 都不可能是查询结果, 证毕。

定义 5(提前终止率) 使用提前终止技术节省的计算代价与不使用提前终止技术的计算代价之比。提前终止率越高, 说明其优化 DTW 计算越有效。

3 DTW 提前终止率的计算方法

3.1 DTW 距离的基本计算方法

计算 DTW 距离时, 通常构造一个 m 行 n 列的累积距离矩阵, 矩阵中每个元素 $\gamma_{i,j}$ 的定义见式(3)。 $\gamma_{i,j}$ 即为序列 $Q[1:j]$ 与序列 $C[1:i]$ 之间的 DTW 距离。

$$\gamma_{i,j} = D_{\text{base}}(q_i, c_j) + \min\{\gamma_{i,j-1}, \gamma_{i-1,j}, \gamma_{i-1,j-1}\} \quad (3)$$

基本计算过程为: 根据式(3), 按照一定的方向逐行(列)计算矩阵元素 $\gamma_{i,j}$, 直至求出 $\gamma_{m,n}$; 并用它与阈值 ϵ 比较: 当 $\gamma_{m,n} \leq \epsilon$ 时, C 是查询结果, 否则 C 不是查询结果。整个过程需要计算累积距离矩阵中全部 $m \times n$ 个元素。

3.2 DTW 距离的提前终止方法

定理 2 若在累积距离矩阵中某一行(列)元素均大于 ϵ , 则有 $D_{\text{dtw}}(Q, C) > \epsilon$ 。

证明见文献[5]。按一定的方向逐行(列)计算累积距离矩阵, 完成一行(列)的计算时, 如果该行(列)上所有元素值都大于 ϵ , 则有效路径溢出, 计算提前终止, 如图 1 所示。

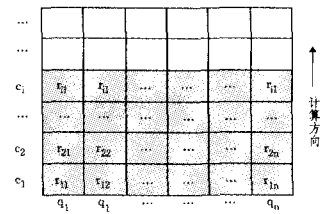


图 1 有效路径溢出

文献[9]在分析 DTW 距离的提前终止方法时, 不用具体计算累积距离矩阵中的某些元素, 只需要给出一个值域区间, 但它涉及比较运算, 同样会带来计算开销, 因此本文也视这些元素参与了距离计算。

3.3 DTW 距离的提前终止率

由上述 DTW 基本计算方法和提前终止方法, 根据定义 5 可得 DTW 的提前终止率, 见式(4)。 N_t, N_c 分别表示一个累积距离矩阵中元素总数和实际计算的元素数量。

$$eff_{\text{real}} = \frac{(N_t - N_c)}{N_t} \quad (4)$$

4 DTW 提前终止率的估算方法

上面给出了 DTW 提前终止率的计算方法。对于一个时间序列数据集, 必须计算两两序列间的 DTW 距离, 才能获得数据集的提前终止率。下面提出一种提前终止率的估算方法, 它选取数据集中的一部分序列作为研究对象, 依统计学原理估算出数据集的平均提前终止率。

4.1 DTW 提前终止率的估算模型

计算 DTW 距离分为两步: 确定序列间的对齐匹配关系, 形成有效路径, 如图 2 所示; 计算有效路径上各点对距离之和^[10], 见式(2)。在求解时, 这两步是通过计算距离累积矩阵实现的, 是同时进行的, 而不是先找出全部匹配关系, 而后计算点对距离之和。假设有效路径上, 每一个点对的计算代价相同(即各点对匹配关系的确定和距离计算的代价分别相同), 则有如下定义。

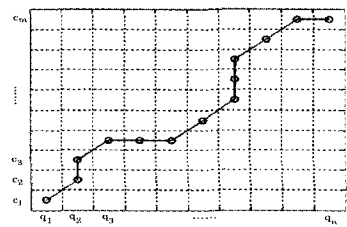


图 2 DTW 距离的有效路径

定义 6(提前终止率估算模型) 使用提前终止技术前后, 有效路径上计算的点对数分别为 c_n 和 c_k , 则提前终止省略了 $c_n - c_k$ 个点对的计算。估算方法见式(5), 模型记为 EaEstDtw。

$$eff_{\text{est}} = (c_n - c_k) / c_n, 0 \leq eff_{\text{est}} < 1 \quad (5)$$

令 Δ_i 为有效路径上第 i 个点对距离, 则

$$\Delta_i = d_i, i = 1, 2, \dots, n \quad (6)$$

C_k 为前 k 个 Δ_i 之和, 即

$$C_k = \sum_{i=1}^k \Delta_i, 1 \leq k \leq n \quad (7)$$

把 Δ_i 视为独立同分布的随机变量, 令 $E(\Delta_i) = \mu, D(\Delta_i) = \sigma^2$ 。 P_{Δ_k} 表示有效路径在 k 处已经溢出的概率, 即

$$P_{\alpha k} = P\{C_k > \epsilon\} \quad (8)$$

P_k 表示有效路径正好在 k 处溢出的概率,可理解为有效路径在 $k-1$ 处没有溢出,而在 k 处恰好溢出的概率。

$$P_k = P\{C_k > \epsilon, C_{k-1} \leq \epsilon\} \quad (9)$$

根据定理 1,发现有效路径溢出就立刻停止计算,则有效路径上需要计算的点对数可用溢出位置的数学期望表示,即

$$E(k) = \sum_{k=1}^n (k \times p_k) \quad (10)$$

当有效路径长度为 c_n 时,提前终止率用式(11)估算。

$$eff_{est} = [c_n - E(k)] / c_n \quad (11)$$

$P_{\alpha k}$ 、 P_k 的计算方法分别由性质 1、性质 2 给出。

性质 1 $P_{\alpha k} \approx 1 - \Phi(\frac{\epsilon - k\mu}{\sqrt{k}\sigma})$, 其中 $\Phi(t)$ 为标准正态分布函数,即 $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{x^2}{2}} dx$ 。

证明:令 $f_k = \frac{C_k - E(C_k)}{\sqrt{D(C_k)}}$, $P_{\alpha k} = P\{C_k > \epsilon\}$, 所以 $P\{C_k > \epsilon\} = P\{\frac{C_k - E(C_k)}{\sqrt{D(C_k)}} > \frac{\epsilon - E(C_k)}{\sqrt{D(C_k)}}\} = P\{f_k > \frac{\epsilon - E(C_k)}{\sqrt{D(C_k)}}\}$ 。

由于 $E(C_k) = k\mu$, $D(C_k) = k\sigma^2$, Δ_k 独立同分布,根据中心极限定理^[11],有 f_k 服从标准正态分布,因此有 $P_{\alpha k} = 1 - P\{f_k \leq \frac{\epsilon - E(C_k)}{\sqrt{D(C_k)}}\} \approx 1 - \Phi(\frac{\epsilon - k\mu}{\sqrt{k}\sigma})$, 证毕。

性质 2 当 $k > 1$ 时, $P_k = P_{\alpha k} - P_{\alpha(k-1)}$; 当 $k=1$ 时, $P_1 = P_{\alpha 1}$ 。

证明:令 A_k 表示有效路径在 k 处溢出的事件, $\overline{A_k}$ 表示有效路径在 k 处没有溢出的事件。当 $k > 1$ 时,由式(9)知, $P_k = P\{C_k > \epsilon, C_{k-1} \leq \epsilon\} = P(A_k \overline{A_{k-1}})$ 。

显然有 $P(A_k) = P(A_k A_{k-1}) + P(A_k \overline{A_{k-1}})$, 其中 $A_k A_{k-1}$ 表示有效路径在 $k-1$ 处溢出的同时在 k 处也溢出的事件。有效路径在 $k-1$ 处溢出,则在 k 处必然溢出,因此可得 $P(A_k A_{k-1}) = P(A_{k-1})$ 。

因为 $P(A_k \overline{A_{k-1}}) = P(A_k) - P(A_{k-1})$, $P(A_k) = P_{\alpha k}$, $P(A_{k-1}) = P_{\alpha(k-1)}$, 所以 $P_k = P_{\alpha k} - P_{\alpha(k-1)}$ 。

当 $k=1$ 时,显然有 $P_1 = P_{\alpha 1}$ 。证毕。

4.2 DTW 提前终止率的估算步骤

估算模型以统计原理为基础,用样本的统计量代替数据集的统计量来估算提前终止率,步骤如下:

- (1) 从时间序列数据集中任选 l_c 个序列作为样本序列;
- (2) 在样本序列中任选查询序列 Q , 顺序计算它与其样本序列的 $l_c - 1$ 条有效路径;
- (3) 在样本序列中,依次以其他 $l_c - 1$ 个序列作为查询序列 Q , 顺序计算它与其样本序列的有效路径;
- (4) 以步骤(2)、(3)中弯曲路径上的点对距离作为样本点,求期望和方差,得到性质 1 中的 μ 和 σ^2 ;
- (5) 以步骤(2)、(3)中有效路径的平均长度作为式(11)中的 c_n ;
- (6) 由 μ 、 σ^2 和 c_n , 根据式(11)估算提前终止率。

5 实验与分析

5.1 实验环境与实验方法

实验环境为 Matlab 7.0, Windows XP Professional SP3,

300G 硬盘, 1.98G 内存, Intel(R) Core(TM)2 Duo CPU。选取 4 组时间序列数据集作为研究对象: 两组为人工合成数据集, 分别由满足标准正态和指数分布的随机数发生器产生, 都由 1000 条长度为 50 的随机序列组成(分别记为 NormDs 和 ExpDs); 另两组为实际的多元时间序列数据集, 分别是 Australian Sign Language^[12](记为 ASL) 和 LP1^[13]。以 ASL 中前 4 种语义对应的序列作为实验数据集, 其包含 22 个变量, 长度在 47~85 之间, 一共 108 个序列; LP1 包含 6 个变量, 长度均为 15, 一共 88 个序列。

实验 1 计算 DTW 提前终止率

实验步骤为: 从容量为 l 的数据集中任取一个序列作为查询序列 Q , 顺序计算它与其他序列的 DTW 距离, 用式(4)计算一次顺序查询的提前终止率; 再以其他序列依次作为查询序列, 重复实验, 可以得到 l 个提前终止率, 数学期望按式(12)确定。提前终止率越大, 表示节省计算的方格数越多, 提前终止策略对 DTW 距离计算的优化程度就越高。

$$E(eff_{real}) = \frac{1}{l} \sum_{i=1}^l eff_{real}(i) \quad (12)$$

实验 2 比较 EaEstDtw、EaEst 方法提前终止率的估算精度

上面介绍了 EaEstDtw 对提前终止率的估算步骤。EaEst 估算方法参见文献[1]。在 NormDs、ExpDs 中选取 10% 的序列作为样本。ASL、LP1 中序列总数不多, 选取 50% 的序列作为样本, 进行提前终止率的估计。以实验 1 计算的提前终止率作为估算方法的比较依据, 估算结果越接近实际的提前终止率, 说明估算精度越高, 估算方法越有效。

实验 3 比较 EaEstDtw 方法与实际求解提前终止率的耗时间

在实验 1、实验 2 的基础上, 分别测量在同等参数条件下, 应用 EaEstDtw 方法和直接计算提前终止率的时间消耗, 多次测量, 结果取平均值。

5.2 实验结果与分析

针对 4 组时间序列数据集, 根据实验 1、实验 2 的方法, 分别获得提前终止率的实际值和估算值, 结果见图 3。图中横坐标表示距离差异阈值 ϵ , 纵坐标表示提前终止率。两种方法的估算误差见表 1。

实验结果表明:

(1) 当距离差异阈值 ϵ 较小时, 提前终止率较高。随着距离差异阈值 ϵ 的增大, 提前终止率减小; 阈值接近实际的 DTW 距离时, 所跳过的方格数急剧减少; 当阈值超过 DTW 距离时, 提前终止率减为 0。

这是由于在 ϵ 超过实际的 DTW 距离时, 累积距离矩阵中没有整行(列)被跳过。但在实际的时间序列搜索中, 选取的 ϵ 比较小, 搜索出来的结果序列一般远小于总的候选集合, 因此提前终止策略使计算资源有更高的利用率。

(2) EaEstDtw 估算结果与实际的提前终止率较为接近, 且比 EaEst 方法有更高的估算精度; EaEstDtw 针对前两组数据集的估算精度优于后两组数据集。

主要原因在于: 前两组数据集由随机序列构成, 有效路径上点对距离更符合“独立同分布随机变量”的假设, 因此估算

(下转第 186 页)

[6] Keogh E, Lin J. Finding unusual medical time-series sub-sequences; algorithms and applications[J]. IEEE Transactions on Information Technology in Biomedicine, 2006, 10(3): 429-439

[7] Keogh E, Lin J, Lee S-H, et al. Finding the most unusual time series subsequence; algorithms and Applications[J]. Knowl Inf Syst, 2006, 11(1): 1-27

[8] 林果园, 郭山清. 基于动态行为和特征模式的异常检测模型[J]. 计算机学报, 2006, 29(9): 1553-1559

[9] 周大钊, 李敏强. 基于序列重要点的时间序列分割[J]. 计算机工程, 2008, 34(27): 14-16

[10] 贾素玲, 陈当阳, 姜浩. 时序数据挖掘中的数据表示算法[J]. 计算机工程与应用, 2006, 29: 184-186

[11] Ameen J R M, Basha R. Hierarchical Data Mining for Unusual Sub-sequence Identifications in Time Series Processes[C]// ICICIC. 2009

[12] 詹艳艳, 陈晓云, 徐荣聪. 基于时间序列模式表示的异常检测算法[J]. 计算机应用, 2007, 24(11): 96-99

[13] Bu Y, Leung T W, Fu A W C, et al. WAT: Finding Top-K Discords in Time Series Database[C]// Proceedings of 7th SIAM. 2007

[14] 喻高瞻, 彭宏, 胡劲松. 时间序列数据的分段线性表示[J]. 计算机应用与软件, 2007, 24(12): 17-18

[15] Ameen J, Basha R. Mining Multiple Periodic Time Series for Detecting Unusual Sub-Sequences[C]// ICICIC. 2009: 1473

(上接第 167 页)

精度优于 ASL、LP1 数据集。

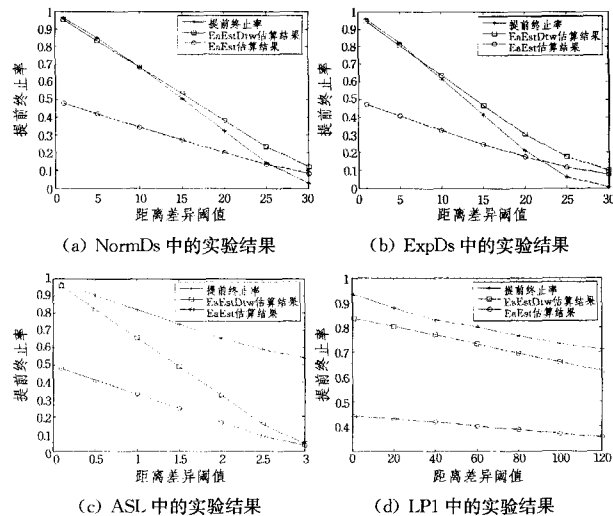


图 3 两种估算方法在 4 组数据集上的实验结果

表 1 EaEstDtw、EaEst 方法的估算误差

数据集	EaEstDtw 方法		EaEst 方法	
	最大绝对误差	平均绝对误差	最大绝对误差	平均绝对误差
NormDs	0.0938	0.0432	0.4854	0.2389
ExpDs	0.1156	0.0580	0.4851	0.2173
ASL	0.4968	0.2523	0.5065	0.4945
LP1	0.0970	0.0759	0.4916	0.4067

针对 NormDs、ExpDs, 用实验 3 描述的方法, 比较 EaEstDtw 与实际求解提前终止率的耗费时间, 结果见图 4。从图中可以看出, EaEstDtw 方法估算提前终止率的计算时间远小于求解实际值。这是因为 EaEstDtw 只需获取一定数目的点对距离作为样本点, 而不用逐个计算数据集中 DTW 累积距离矩阵, 因此比求解实际值节省了大量计算。

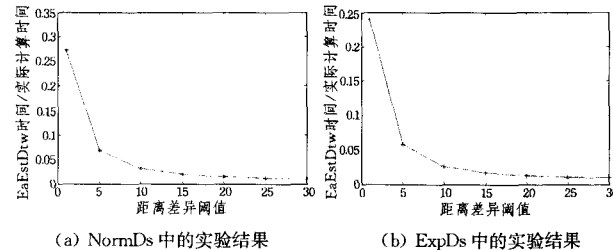


图 4 EaEstDtw 与实际求解提前终止率耗费时间的比较

结束语 本文分析了 DTW 提前终止产生的机理, 提出了一种 DTW 提前终止率的估算模型, 并通过实验验证了方

法的有效性。实验结果表明, EaEstDtw 能够以相对较小的计算代价, 有效地估算 DTW 提前终止率; 估算精度明显优于 EaEst 方法; EaEstDtw 方法对随机序列有着较高的估算精度, 在一定程度上也适用于实际产生的多元时间序列数据集。

参考文献

[1] 李俊奎, 王元珍, 李海波, 等. 一种时间序列相似搜索中提前终止效率的估算方法[J]. 计算机科学, 2009, 36(1): 114-117

[2] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[C]// Proc of the KDD Workshop. Seattle, WA, 1994: 359-370

[3] Vlachos M, Hadjieleftheriou M, Gunopulos D, et al. Indexing multi-dimensional time-series with support for multiple distance measures[C]// Proc of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington, 2003: 216-255

[4] Keogh E J. Exact indexing of dynamic time warping[C]// Proceedings of 28th International Conference on Very Large Data Bases. Hong Kong, China, 2002: 406-417

[5] Kim M-S, Kim S-W, Shin M. Optimization of Subsequence Matching Under Time Warping in Time-series Databases[C]// ACM Symposium on Applied Computing. New Mexico, USA, 2005: 581-586

[6] Li Jun-kui, Wang Yuan-zhen. EA-DTW: Early Abandon to Accelerate Exactly Warping Matching of Time Series[C]// Proc of Int'l Conf. on Intelligent Systems and Knowledge Engineering. 2007

[7] 陈胜利, 李俊奎, 刘小东. 基于提前终止的加速时间序列弯曲算法[J]. 计算机应用, 2010, 30(4): 1068-1071

[8] 李爱国, 覃征. 大规模时间序列数据库降维及相似搜索[J]. 计算机学报, 2005, 28(9): 1467-1475

[9] 李俊奎. 时间序列相似性问题研究[D]. 武汉: 华中科技大学计算机科学与技术学院, 2008

[10] 穆斌, 闫金来. 高效的时间序列下界技术[J]. 计算机工程与应用, 2009, 45(11): 168-171

[11] 盛骤, 谢式千, 潘承毅. 概率论与数理统计[M]. 北京: 高等教育出版社, 2001

[12] [http://kdd.ics.uci.edu/databases/High-quality Australian Sign Language/High-quality Australian Sign Language. html](http://kdd.ics.uci.edu/databases/High-quality%20Australian%20Sign%20Language/High-quality%20Australian%20Sign%20Language.html)

[13] [http://kdd.ics.uci.edu/databases/robotfailure/robotfailure. html](http://kdd.ics.uci.edu/databases/robotfailure/robotfailure.html)