软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(3):560−575 [doi: 10.13328/j.cnki.jos.004410]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种支持 DTW 距离的多元时间序列索引结构*

李正欣, 张凤鸣, 李克武, 张晓丰

(空军工程大学 装备管理与安全工程学院,陕西 西安 710051)

通讯作者: 李正欣, E-mail: lizhengxin_2005@163.com

摘 要: 现有的索引结构难以有效地支持 DTW 距离度量下的多元时间序列相似性搜索.首先给出一种将不等长多元时间序列转换为等长一元时间序列的方法,并证明这种转换满足下界距离引理;以此为基础,提出一种多元时间序列的 DTW 下界距离,并对其性质进行分析;然后,针对给出的下界距离,提出一种支持 DTW 距离度量的多元时间序列索引结构,对多元时间序列数据库进行有效组织;再给出多元时间序列相似模式搜索算法及流程,并证明该搜索方法具有非漏报性;最后,通过实验对所提方法的有效性进行验证.

关键词: 多元时间序列;动态时间弯曲;下界距离;索引结构;相似性搜索

中图法分类号: TP311 **文献标识码**: A

中文引用格式: 李正欣,张凤鸣,李克武,张晓丰.一种支持 DTW 距离的多元时间序列索引结构.软件学报,2014,25(3):560−575. http://www.jos.org.cn/1000-9825/4410.htm

英文引用格式: Li ZX, Zhang FM, Li KW, Zhang XF. Index structure for multivariate time series under DTW distance metric. Ruan Jian Xue Bao/Journal of Software, 2014,25(3):560−575 (in Chinese). http://www.jos.org.cn/1000-9825/4410.htm

Index Structure for Multivariate Time Series under DTW Distance Metric

LI Zheng-Xin, ZHANG Feng-Ming, LI Ke-Wu, ZHANG Xiao-Feng

(Equipment Management and Safety Engineering College, Air Force Engineering University, Xi'an 710051, China)

Corresponding author: LI Zheng-Xin, E-mail: lizhengxin_2005@163.com

Abstract: Existing index structures for multivariate time series can't support similarity search under DTW distance efficiently. Firstly, a transformation method, which converts unequal-length multivariate time series into equal-length univariate time series, is proposed and a mathematical proof that the transformation satisfies lower bounding distance lemma is provided. Secondly, DTW lower bounding distance is proposed, and its character is analyzed. Thirdly, based on DTW lower bounding distance proposed above, an index structure for multivariate time series is proposed, allowing database of multivariate time series be organized. Further, similarity search algorithm and process for multivariate time series are discussed, and related mathematical proofs that false dismissals can be avoided are given. Finally, validity of proposed method is verified by experiments.

Key words: multivariate time series; dynamic time warping; lower bounding distance; index structure; similarity search

现实世界中存在大量多元时间序列(multivariate time series,简称 MTS)类型的数据[1],如航天飞船等重要仪器的运行状态数据、互联网中关键服务器的通信流量数据、应用于多种行业的人体运动捕捉数据、患者的脑电波等.一些多媒体数据经过转换后也可以形成多元时间序列.时间序列相似模式挖掘就是从时间序列数据库中查找和发现用户感兴趣的模式,旨在研究隐含在时间序列之中的更深层次的知识,从中获取蕴含的系统演化规律[2].相似模式挖掘不仅是时间序列数据挖掘的主要研究内容之一,还是实现其他挖掘任务的基础[3].

模式匹配与相似性搜索是时间序列相似模式挖掘的两个核心问题.模式匹配是度量时间序列相似程度的方法,在时间序列分析处理中具有基础性地位[4].目前,最常见的 MTS 模式匹配方法是 Minkowski 距离、动态时间弯曲(dynamic time warping,简称 DTW)距离.Minkowski 距离计算简单、容易理解,但它要求两条时间序列的

* 收稿时间: 2011-10-08; 修改时间: 2012-10-30; 定稿时间: 2013-04-02

长度必须相等,且对时间轴伸缩和弯曲问题无能为力.DTW 距离定义了序列之间的最佳对齐匹配关系,支持不同长度时间序列的相似性度量,支持时间轴的伸缩和弯曲[5].由于 DTW 距离比 Minkowski 距离有更好的鲁棒性,因此被广泛用于时间序列的相似性度量[6].在时间序列相似性搜索中,如果用查询序列逐一与数据库中其他序列进行相似性比较,搜索效率很低,在实际应用中往往是不可行的.因此,研究高效的搜索方法是非常必要的.目前,针对 Minkowski 距离度量的索引方法较为成熟[7],而针对 DTW 距离度量的索引方法并不多见.主要原因是 DTW 距离不满足距离三角不等式,且计算复杂度较高[8].

已有的支持 DTW 距离度量的索引结构基本都遵循如下思路:寻找一种计算更简单的距离度量来粗略地估计 DTW 距离,称为 DTW 下界距离,通过它过滤掉大部分不满足相似性要求的序列,从而提高查询效率.为了保证查询的准确和高效,DTW 下界距离应满足 3 个条件[9]:

- 正确性:经下界距离过滤得到的候选集中必须包含所有满足条件的序列,即不允许出现漏报;
- 有效性:下界距离的计算复杂度应尽量低;
- 紧致性:下界距离的度量结果应尽量逼近 DTW 距离,这样才能使得候选集不至过大,从而减少后处理的计算量.

Yi[10],Kim[11],Keogh[12]和 Zhu[13]分别提出了支持 DTW 距离度量的一元时间列搜索方法,他们分别给出了各自的 DTW 下界距离,然后提出了支持相应下界距离的索引构建方法,并且证明搜索方法的非漏报性.Yi 计算两条一元时间序列的 DTW 下界距离时,选择一条序列作为基准序列,以另一条序列中大于基准序列最大值的点集以及小于基准序列最小值的点集作为特征,以此为基础构造 DTW 下界距离,记为 *LB_Yi*.Kim 提取一元时间序列的起始点、结束点、最大值点和最小值点这 4 个特征,以此为基础构造 DTW 下界距离,记为 *LB_Kim*.Keogh 提取查询序列的上、下边界序列作为查询特征,进而构造出一种 DTW 下界距离,记为 *LB_Keogh*;使用 PAA 方法把数据库中的一元时间序列转换为空间向量点,用 R-Tree 对向量点进行组织;利用下界距离 *LB_Keogh* 在空间索引结构上执行查询,索引查询的结果构成候选集;最后,使用 DTW 距离计算查询序列与候选集中每个一元时间序列的 DTW 距离,去除不符合相似性条件的序列,得到结果集,并通过大量实验验证了 *LB_Keogh* 的紧致性优于 *LB_Yi* 和 *LB_Kim*.Zhu 在文献[13]中对下界距离方法进行了数学证明,并提出了一种 DTW 下界距离,记为 *LB_Zhu*,这可以视为 *LB_Keogh* 方法的改进,进一步提高了下界距离在索引查询中的紧致性.此外,文献[14]对时间序列进行分段累积近似,用网格最小边界矩形近似表示查询序列,进而提出一种 DTW 下界距离,记为 *LB_GMBR*.

文献[10–14]提出的搜索方法不会遗漏正确结果,下界距离 *LB_Keogh* 及其改进形式 *LB_Zhu* 的紧致性较高,相应搜索方法的整体性能较优.但以上几种方法存在一定的局限性:它们仅针对一元时间序列,而不适用于多元时间序列;Keogh 和 Zhu 提出的方法还要求查询序列和搜索序列的长度必须相等.

本文的目标是找到一种支持 DTW 距离度量的多元时间序列索引结构,从而实现多元时间序列的高效搜索.首先给出一种多元时间序列的 DTW 下界距离;在下界距离的基础上,提出一种支持 DTW 距离度量的多元时间序列索引结构,进而给出相应的相似性搜索算法;最后,通过实验对所提方法进行有效性分析.

1 预备知识

1.1 DTW距离及性质分析

定义 1(时间序列). 一系列记录值 $x_t(j)$称为时间序列(time series,简称 TS),其中,$t(t=1,2,\ldots,n)$表示第 t 个时间点,$j(j=1,2,\ldots,m)$表示第 j 个变量,$x_t(j)$表示第 j 个变量在第 t 个时间点上的记录值.当 $m>1$ 时,$x_t(j)$为多元时间序列;当 $m=1$ 时,$x_t(j)$为一元时间序列(univariate time series,简称 UTS).时间序列可以用 $m\times n$ 矩阵表示,m 表示变量数,n 表示时间点数量,矩阵行代表变量维,列代表时间维.

定义 2(DTW 距离)[15]. 设时间序列 $X=\langle x_1,x_2,\ldots,x_n\rangle$,$Y=\langle y_1,y_2,\ldots,y_m\rangle$,则 X,Y 的 DTW 距离 $D_{dtw}(X,Y)$定义见公式(1),其中,$D_{base}(x_i,y_j)$表示向量点 x_i 和 y_j 之间的基距离,可以根据情况选择不同的距离度量.不失一般性,本文使用 Minkowski 距离作为基距离.

$$D_{dtw}(X,Y)=D_{base}(x_1,y_1)+\min\begin{cases}D_{dtw}(X,Y[2:-])\\D_{dtw}(X[2:-],Y)\\D_{dtw}(X[2:-],Y[2:-])\end{cases} \tag{1}$$

DTW 距离允许序列点自我复制后再进行对齐匹配,能够很好地支持时间轴弯曲,并且它可以对非等长时间序列进行度量,也支持时间轴伸缩,但其计算复杂度较高、且不满足距离三角不等式.DTW 距离实际上就是确定序列 X 与 Y 上每个点之间的对齐匹配关系,如图 1(c)所示,这种匹配关系可能有很多种,每一种匹配关系可以用一条弯曲路径表示,如图 1(b)所示.也就是说,序列间的匹配关系与弯曲路径是一一对应的关系.

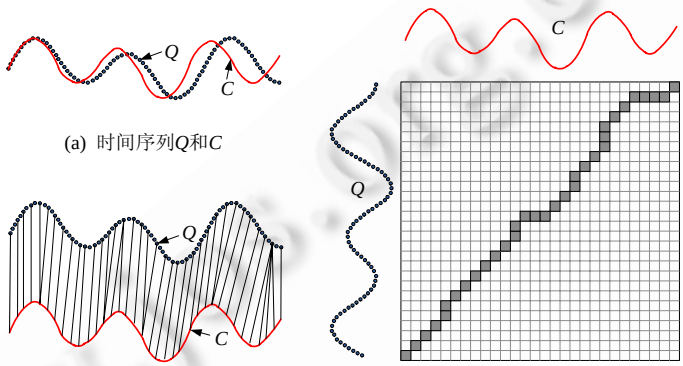


Fig.1 Warping path and resulting alignment

图 1 弯曲路径与点对匹配结果

弯曲路径必须满足 3 个基本条件:

(1) 边界条件:路径起始于点(x_1,y_1)、终止于点(x_n,y_m),它表示两个序列的起始点和结束点对应匹配;

(2) 连续性:路径上的任意两个相邻点 (x_{i_1},y_{j_1}) 和 (x_{i_2},y_{j_2}) 满足条件 $0\leqslant|i_1-i_2|\leqslant1,0\leqslant|j_1-j_2|\leqslant1$;

(3) 单调性:若 (x_{i_1},y_{j_1}) 和 (x_{i_2},y_{j_2}) 为路径上前后两个点,则须满足 $i_2-i_1\geqslant0,j_2-j_1\geqslant0$.

满足上述条件的弯曲路径有很多,每一条弯曲路径都代表一种点对匹配关系.设弯曲路径为 $W=(w_1,w_2,\ldots,w_k,\ldots,w_K),w_k=(i,j)_k$ 是弯曲路径上第 k 个元素,它表示 x_i 与 y_j 建立的匹配关系,路径长度满足 $\max(n,m)\leqslant K\leqslant n+m-1$.

点对匹配关系中,点对基距离之和的最小值即为 DTW 距离,对应的弯曲路径为最佳路径.DTW 距离表示为

$$DTW(X,Y)=\min\left\{\sum_{k=1}^{K}D_{base}(w_k)\right\} \tag{2}$$

求解最佳路径需要构造一个 m 行 n 列的累积距离矩阵 $M_{m\times n}$,矩阵中的每个元素 $\gamma_{i,j}$ 定义为

$$\gamma_{i,j}=D_{base}(x_i,y_j)+\min\{\gamma_{i,j-1},\gamma_{i-1,j},\gamma_{i-1,j-1}\} \tag{3}$$

$\gamma_{i,j}$ 为序列 $X[1:j]$与序列 $Y[1:i]$的 DTW 距离,因此,$D_{dtw}(X,Y)=\gamma_{m,n},\gamma_{m,n}$ 可以用动态规划法求解[16].

1.2 查询策略与查询完备性

顺序扫描是用查询序列逐一与 MTS 数据库中的序列进行模式匹配,找出满足相似条件的序列.然而,数据库包含的序列数量很多,且 DTW 距离的计算复杂度较高,因此,顺序扫描方法效率很低.目前常用的查询策略一般遵循两步查询步骤(如图 2 所示):

(1) 先使用下界距离进行索引查询:将时间序列映射到低维特征空间,转换为低维空间中的几何对象,采用空间索引结构组织这些低维空间对象;然后,使用查询序列的特征在索引结构上进行查询,通过索引的过滤和剪枝策略提高查询效率,索引查询结果即为候选集;

(2) 再使用 DTW 距离对候选集进行后处理:依次计算查询序列与候选集中每个序列的 DTW 距离,去除

不符合相似性条件的序列,得到结果集.

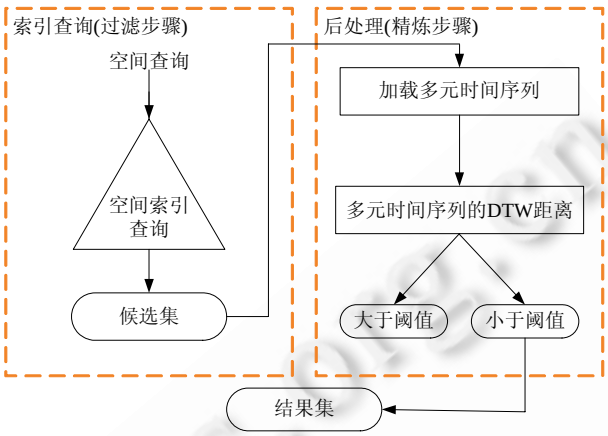


Fig.2 Process of two-step search

图 2 两步搜索流程

查询完备性是衡量查询策略优劣的一个重要标准,它包括两层含义:完全性和准确性.设 S 是数据库 TB 中满足相似性匹配要求的序列集合,R 是实际查询到的序列集合.若 R 是 S 的子集,则查询不是完全的,$S-R$ 表示遗漏的正确结果,称发生漏报(false dismissal);相反地,若 S 是 R 的子集,则查找是不准确的,$R-S$ 表示引入的错误结果,称发生误报(false alarm).

通常,查询准确性较容易得到保证,只需结果集中的序列都满足相似模式匹配要求;而查询完全性却并不是所有的查询策略都能达到的,有时会为了查询的效率而损失一定的查询完全性.Faloutsos 等人在文献[17]中证明了一个重要引理,能够保证时间序列在变换到特征空间后的相似模式搜索不发生漏报.

定理 1(下界距离引理). 设时间序列 Q 和 C 通过特征提取函数 F 映射到特征空间,为了保证特征空间的搜索不产生漏报,必须满足 $D_{feature}(F(Q),F(C))\leqslant D_{\text{true}}(Q,C)$,其中,$D_{feature}$ 和 D_{true} 分别表示特征空间和原始空间的距离度量函数.

1.3 R-Tree索引结构

R-Tree 最初由 Guttman 于 1984 年提出,随后人们在此基础上针对不同的空间操作需求提出了各种改进方案,如 R^+-Tree,R^*-Tree 等,经过 20 多年的发展,逐渐形成了一个枝繁叶茂的空间索引 R-Tree 家族[18].R-Tree 是一种处理多维数据的空间索引结构,是许多空间索引方法的基础,在空间索引领域中占有重要地位[19].它的结点分为两类:内部结点和叶结点.内部结点包括若干个形如(ptr,R)的项,其中,ptr 是指向树中下一层结点的指针,R 是包括 ptr 所指向结点中的所有最小界限矩形(minimum bounding rectangle,简称 MBR)的最小矩形.叶结点包括若干个形如(oid,R)的项,其中,oid 是指向目标对象的标识符,R 是目标对象的 MBR.

2 多元时间序列的 DTW 下界距离

本节首先给出一种将不等长 MTS 转换为等长 UTS 的方法,并证明这种转换满足下界距离引理.以此为基础,提出一种多元时间序列的 DTW 下界距离,并对其性质进行分析.

2.1 弯曲路径的全局约束条件

除了定义的 3 个条件之外,弯曲路径还需满足全局约束条件,即限定一个序列中的点只能同另一序列中位置相近的某些点进行匹配,累积距离矩阵中允许弯曲路径访问的元素集合被称为弯曲窗口.设时间序列 Q,C 的弯曲路径上的元素为 $w_k=(i,j)_k$,弯曲路径的全局约束条件可以理解为对元素 $w_k=(i,j)_k$ 下标的限制,即 $j-r\leqslant i\leqslant j+r$,

其中,r 表示序列上某个点的弯曲限制.Sakoe-Chiba 约束中,r 为常数,弯曲窗口为沿对角线方向的带形,如图 3(a)所示;Itakura-Parallelogram 约束中,r 为 i 的函数,弯曲窗口为沿对角线方向的平行四边形,如图 3(b)所示.文中主要针对 Sakoe-Chiba 约束条件进行讨论.

(a) Sakoe-Chiba 弯曲窗口 (b) Itakura-Parallelogram 弯曲窗口

Fig.3 Two kinds of warping windows

图 3 两种弯曲窗口

性质 1. 设时间序列 $Q=\langle q_1,q_2,\ldots,q_m\rangle$,$C=\langle c_1,c_2,\ldots,c_n\rangle$,弯曲路径在全局约束条件下的弯曲限制为 r,为了保证 DTW 距离计算的有效性,Q,C 的长度差不大于 r.

证明:使用反证法.

设时间序列 Q,C 的长度分别为 m,n,它们的长度差大于 r,即 $m-n>r$ 或 $m-n<-r$.

DTW 距离计算的有效性可以理解为:累积距离矩阵中,弯曲路径除了要满足自身定义的 3 个条件之外,还必须满足全局约束条件.

根据第 1 个约束条件,弯曲路径的起始点由 q_1,c_1 形成,终止点由 q_m,c_n 形成;如果点 q_i,c_j 是弯曲路径上的匹配点对,其中,$1\leqslant i\leqslant m$,$1\leqslant j\leqslant n$,则全局约束条件要求 $j-r\leqslant i\leqslant j+r$.

由于 q_m,c_n 形成弯曲路径的终止点,因此满足 $n-r\leqslant m\leqslant n+r$,显然与假设矛盾. □

性质 1 表明:计算序列 Q,C 的 DTW 距离时,当增加弯曲路径的全局约束条件后,仍然支持不等长序列的匹配,但序列的长度差有一定限制.本文余下部分都是在性质 1 的条件下进行讨论.

2.2 不等长MTS转换为等长UTS

MTS 在时间维和变量维上都具有较高的维度数,直接用索引结构对其进行组织较为困难,为了便于用索引结构对 MTS 进行有效组织,文中将不等长 MTS 转换为等长 UTS.下面提出几个性质,并对其进行简单证明,作为这种变换的理论基础.

性质 2. 设时间序列 $Q=\langle q_1,q_2,\ldots,q_m\rangle$,$C=\langle c_1,c_2,\ldots,c_n\rangle$,在累积距离矩阵 M 中,如果存在一条弯曲路径,该路径上的点对基距离之和为 α,则 $DTW(Q,C)\leqslant\alpha$.

证明:在累积距离矩阵中,能够找到多条弯曲路径,其中存在一条最佳路径 $W=(w_1,w_2,\ldots,w_k,\ldots,w_K)$,元素 $w_k=(i,j)_k$ 表示 q_i 与 c_j 建立的匹配关系,使得路径 W 确定的点对基距离之和 $\sum_{i=1}^{K}D_{base}(w_k)$ 最小,记

$$D_{base}(w_k)=D_{base}(q_i,c_j).$$

现存在一条弯曲路径 $W'=(w'_1,w'_2,\ldots,w'_k,\ldots,w'_{K'})$,使得 $\sum_{k=1}^{K'}D_{base}(w'_k)=\alpha$.

显然有 $\sum_{k=1}^{K}D_{base}(w_k)\leqslant\sum_{k=1}^{K'}D_{base}(w'_k)$.

由 DTW 距离的定义,$DTW(Q,C)=\sum_{k=1}^{K}D_{base}(w_k)$,因此 $DTW(Q,C)\leqslant\alpha$. □

性质 3. 设时间序列 $Q=\langle q_1,q_2,\ldots,q_m\rangle$,$Q'=\langle q'_1,q'_2,\ldots,q'_m\rangle$,$C=\langle c_1,c_2,\ldots,c_n\rangle$,$C'=\langle c'_1,c'_2,\ldots,c'_n\rangle$,如果 Q' 与 Q、C' 与 C 的长度分别相同,且 Q',C' 上任意点对的基距离均不大于 Q,C 上对应点对的基距离,即 $D_{base}(q'_i,c'_j)\leqslant D_{base}(q_i,c_j)$,其中,$1\leqslant i\leqslant m$,$1\leqslant j\leqslant n$,则 $DTW(Q',C')\leqslant DTW(Q,C)$.

证明:首先求解 $DTW(Q,C)$,该过程可理解为两个步骤:确定 Q 与 C 上的点对的最佳匹配关系,形成最佳路径 $W=(w_1,w_2,\ldots,w_k,\ldots,w_K)$,元素 $w_k=(i,j)_k$ 表示 q_i 与 c_j 建立的匹配关系;计算所有匹配点对的基距离之和,即

$$DTW(Q,C)=\sum_{k=1}^{K}D_{base}(w_k).$$

然后求解 $DTW(Q',C')$,由于序列 Q' 与 Q、C' 与 C 的长度分别相同,求解 $DTW(Q',C')$ 时,可以沿用弯曲路径 W 形成的点对匹配关系,并计算这种匹配关系下的点对基距离之和 α.

因为 $D_{base}(q_i',c_j')\leqslant D_{base}(q_i,c_j)$,所以 $\alpha\leqslant\sum_{i=1}^{K}D_{base}(w_k)$,即 $\alpha\leqslant DTW(Q,C)$.

因为在 Q',C' 形成的累积距离矩阵中存在一条弯曲路径,该路径上的点对基距离之和为 α,由性质 2 可知,

$$DTW(Q',C')\leqslant\alpha.$$

因此,$DTW(Q',C')\leqslant DTW(Q,C)$. □

性质 4. 变量维数为 $m(m>1)$的时间序列 Q,C,把任意对应的 $k(1\leqslant k\leqslant m)$个变量相加,分别得到两组一元时间序列 Q',C',则 $DTW(Q',C')\leqslant DTW(Q,C)$.

证明:不失一般性,假设 Q',C' 分别是由 Q,C 的前 $k(1\leqslant k\leqslant m)$个变量相加得到的一元时间序列.显然,$Q'$ 与 Q、C' 与 C 的长度分别相同.

设 Q,C 任意时刻的记录值分别为 $q_i=(q_{i1},q_{i2},\ldots,q_{im})^T$,$c_j=(c_{j1},c_{j2},\ldots,c_{jm})^T$,其中,$1\leqslant i\leqslant Len(Q)$,$1\leqslant j\leqslant Len(C)$;

Q',C' 在对应时刻的值分别为 $q_i'=\sum_{x=1}^{k}q_{ix}$,$c_j'=\sum_{x=1}^{k}c_{jx}$.

Q,C 上任意点对的基距离为 $D_{base}(q_i,c_j)=\sum_{x=1}^{m}|q_{ix}-c_{jx}|$;

Q',C' 上对应点对的基距离为 $D_{base}(q_i',c_j')=|\sum_{x=1}^{k}(q_{ix}-c_{jx})|$.

由于 $\sum_{x=1}^{m}|q_{ix}-c_{jx}|=\sum_{x=1}^{k}|q_{ix}-c_{jx}|+\sum_{x=k+1}^{m}|q_{ix}-c_{jx}|$,$\sum_{x=1}^{k}|q_{ix}-c_{jx}|\geqslant|\sum_{x=1}^{k}(q_{ix}-c_{jx})|$,且

$$\sum_{x=k+1}^{m}|q_{ix}-c_{jx}|\geqslant 0,$$

所以 $\sum_{x=1}^{m}|q_{ix}-c_{jx}|\geqslant|\sum_{x=1}^{k}(q_{ix}-c_{jx})|$.

因此,$D_{base}(q_i',c_j')\leqslant D_{base}(q_i,c_j)$.根据性质 3 可知,$DTW(Q',C')\leqslant DTW(Q,C)$. □

根据性质 4,可以把 MTS 转换为 UTS,文中把这种转换称为变量加和,并且这种转换满足下界距离引理.下面提出一种序列扩展方法,把不等长 UTS 转换为等长 UTS.

设 UTS 数据库中序列的最大长度为 $L\max$,其中,任意两条一元时间序列记为 $Q'=\langle q_1',q_2',\ldots,q_m'\rangle$,$C'=\langle c_1',c_2',\ldots,c_n'\rangle$,长度分别为 m,n.

序列扩展可以表示为映射:$F(Q')\rightarrow Q'^{+}$,F 把任意长度序列映射为长度为 $L\max+1$ 的序列.

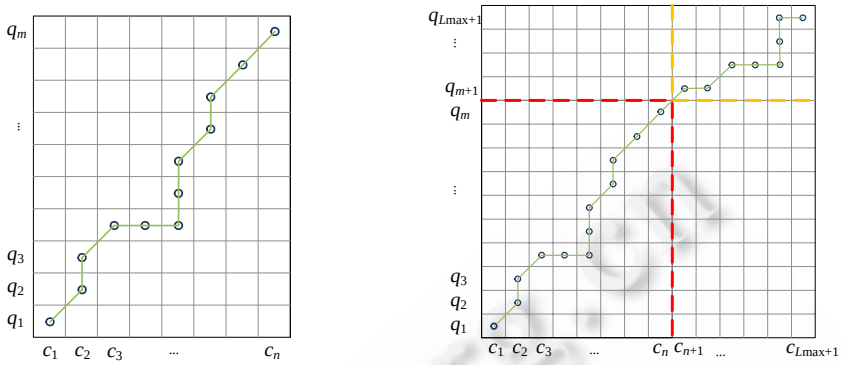
序列扩展方法可描述为:在长度为 L 的原始序列后面增加 $L\max+1-L$ 个常数 e(取 $e=0$).

例如,$Q'^{+}=\langle Q',Q'^{*}\rangle$,$C'^{+}=\langle C',C'^{*}\rangle$,其中,$Q'^{*}=\langle 0,0,\ldots,0\rangle_{L\max+1-m}$,$C'^{*}=\langle 0,0,\ldots,0\rangle_{L\max+1-n}$.

下面简单说明 $DTW(Q'^{+},C'^{+})\leqslant DTW(Q',C')$成立.

Q',C' 的最佳路径 W 如图 4(a)所示,最佳路径确定的匹配点对基距离之和即为 $DTW(Q',C')$.

Q'^{+} 由 Q',Q'^{*} 组成,C'^{+} 由 C',C'^{*} 组成,因此,寻找 Q'^{+},C'^{+} 之间的一种点对匹配关系可分为两个步骤:先确定 Q',C' 上的点对匹配关系;再确定 Q'^{*},C'^{*} 上的点对匹配关系.Q',C' 仍可沿用路径 W,Q'^{*},C'^{*} 任意确定一种匹配关系,如图 4(b)所示.由于 Q'^{*},C'^{*} 上的值都为 0,所以在构造的点对匹配关系下,Q'^{+},C'^{+} 间匹配点对的基距离之和为 $DTW(Q',C')$.即在 Q'^{+}、C'^{+} 形成的累积距离矩阵中存在一条弯曲路径,该路径上的点对基距离之和为 $DTW(Q',C')$,由性质 2 可知,$DTW(Q'^{+},C'^{+})\leqslant DTW(Q',C')$.



(a) Q',C'的最佳路径 (b) Q'^{+},C'^{+}的一条弯曲路径

Fig.4 DTW distance computation of extended time series

图 4 扩展时间序列的 DTW 距离计算

2.3 DTW下界距离

上文提出了把不等长 MTS 转换为等长 UTS 的方法,下面给出等长 UTS 的 DTW 下界距离.

设一元查询序列 $Q=\langle q_1,q_2,\ldots,q_n\rangle$,弯曲路径在全局约束条件下的弯曲限制为 r,定义两条新序列 $U=\langle u_1,u_2,\ldots,u_n\rangle$,$L=\langle l_1,l_2,\ldots,l_n\rangle$:

$$u_i=\max(q_{i-r},q_{i+r}) \tag{4}$$

$$l_i=\min(q_{i-r},q_{i+r}) \tag{5}$$

U,L 分别称为 Q 的上、下边界序列.图 5 反映了 Q 与上、下边界序列的位置关系,Q 被包围在上、下边界序列形成的区域中,该区域称为封袋.显然,有公式(6)成立.

$$\forall i,u_i\geqslant q_i\geqslant l_i \tag{6}$$

等长一元时间序列 Q,C 的 DTW 下界距离定义为:

$$LB\_DTW(U,L,C)=\sum_{i=1}^{n}\begin{cases}|c_i-u_i|, & \text{if } c_i>u_i\\ |c_i-l_i|, & \text{if } c_i<l_i\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

$LB_DTW(U,L,C)$可理解为:C 没有落入封袋的点同封袋边界的距离之和,如图 6 所示.

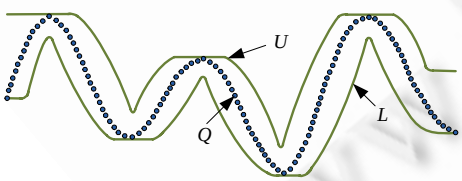


Fig.5 Query series and its upper and lower bounding series

图 5 查询序列与其上、下边界序列

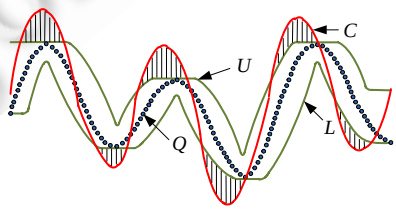


Fig.6 An illustration of the lower bounding function LB_DTW

图 6 下界距离 LB_DTW 的示意图

下面证明 $LB\_DTW(U,L,C)\leqslant LB\_DTW(Q,C)$,即 $LB_DTW(U,L,C)$满足下界距离引理.

证明:设 Q,C 的最佳路径 $W=(w_1,w_2,\ldots,w_k,\ldots,w_K)$,由 DTW 距离的定义得 $DTW(Q,C)=\sum_{k=1}^{K}D_{base}(w_k)$,其中,$n\leqslant K<2n-1$,则原命题可转换为

$$\sum_{i=1}^{n}\begin{cases}|c_i-u_i|, & \text{if } c_i>u_i\\ |c_i-l_i|, & \text{if } c_i<l_i\\ 0, & \text{otherwise}\end{cases} \leqslant \sum_{k=1}^{K} D_{base}(w_k) \tag{8}$$

弯曲路径上的元素 $w_k=(i,j)_k$,每一个 $i(1\leqslant i\leqslant n)$对应一个或多个 j,选出其中 j 值最小的元素,把相应的 k 标记为 $k\in match$.公式(8)变为

$$\sum_{i=1}^{n}\begin{cases}|c_i-u_i|, & \text{if } c_i>u_i\\ |c_i-l_i|, & \text{if } c_i<l_i\\ 0, & \text{otherwise}\end{cases} \leqslant \sum_{\kappa\in match} D_{base}(w_k)+\sum_{\kappa\in unmatch} D_{base}(w_k) \tag{9}$$

公式(9)中,不等式左侧分为 3 种情况:

- 当 $c_i>u_i$ 时,不等式左侧的第 i 项为$|c_i>u_i|$,不等式右侧 $\sum_{\kappa\in match} D_{base}(w_k)$ 中的对应元素为 $D_{base}(w_k)=|c_i-q_j|$.

  因为 $u_i=\max(q_{i-r}:q_{i+r}),i-r\leqslant j\leqslant i+r$,所以 $q_j\leqslant\max(q_{i-r}:q_{i+r})$,即 $q_j\leqslant u_i$;变形后有 $c_i-u_i\leqslant c_i-q_j$,即$|c_i-u_i|\leqslant|c_i-q_j|$,因此,$|c_i-u_i|\leqslant D_{base}(w_k)$;
- 当 $c_i<l_i$ 时,同理可证$|c_i-l_i|\leqslant D_{base}(w_k)$;
- 当 $l_i\leqslant c_i\leqslant u_i$ 时,显然有 $0\leqslant|c_i-q_j|$.

所以有,

$$\sum_{i=1}^{n}\begin{cases}|c_i-u_i|, & \text{if } c_i>u_i\\ |c_i-l_i|, & \text{if } c_i<l_i\\ 0, & \text{otherwise}\end{cases} \leqslant \sum_{\kappa\in match} D_{base}(w_k) \tag{10}$$

弯曲路径上基距离非负,有 $\sum_{\kappa\in unmatch} D_{base}(w_k)\geqslant 0$,因此公式(9)成立. □

下面分别从正确性、有效性和紧致性这 3 个方面对 *LB_DTW* 进行分析:

(1) 设变量维数为 $m(m>1)$的不等长多元时间序列 Q,C,把任意对应的 $k(1\leqslant k\leqslant m)$个变量相加,得到不等长一元时间序列 Q',C',由性质 4 可知,$DTW(Q',C')\leqslant DTW(Q,C)$;再使用序列扩展方法,把不等长一元时间序列 Q',C'扩展为等长一元时间序列 Q'^{+},C'^{+},且有 $DTW(Q'^{+},C'^{+})\leqslant DTW(Q',C')$;对于等长一元时间序列 Q'^{+},C'^{+},设 Q'^{+}的上、下边界序列分别为 U'^{+},L'^{+},则有 $LB\_DTW(U'^{+},L'^{+},C'^{+})\leqslant DTW(Q'^{+},C'^{+})$,因此有 $LB\_DTW(U'^{+},L'^{+},C'^{+})\leqslant DTW(Q,C)$,再根据下界距离引理得知,$LB_DTW(U'^{+},L'^{+},C'^{+})$是 $DTW(Q,C)$的下界距离,用其作为距离度量时,查询结果不会产生漏报;

(2) 从公式(7)可以看出,*LB_DTW* 下界距离是针对等长 UTS 的模式匹配方法,点对匹配关系明确,在形式和计算方法上都非常类似于 Minkowski 距离,因此同 DTW 距离相比,计算复杂度明显降低;

(3) *LB_DTW* 下界距离的紧致性不易定性分析,下文将通过实验对其进行验证.

LB_Keogh 及其改进方法只适用于等长 UTS,具有较大的局限性;而 *LB_DTW* 以性质 1~性质 4 为理论支撑,实现了从一元向多元、从等长向不等长的拓展,把研究对象从等长 UTS 推广到不等长 MTS,可视为 *LB_Keogh* 方法的拓展.

3 支持 DTW 距离的索引结构

针对上文给出的下界距离,本节提出一种支持 DTW 距离的多元时间序列索引结构,对 MTS 数据库进行有效组织,并给出 MTS 相似性搜索算法.

3.1 时间序列的分段累积近似

为了把长度从 n 降到 N,UTS 在时间维度上被分割为等长度的 N 段,用每一段记录值的平均值作为该段序列的基本特征,这种表示方法被称为时间序列的分段累积近似(piecewise aggregate approximation,简称 PAA).

令 N 表示时间序列的分段数目,则序列 C 可用 N 维空间中的点表示为 $\overline{C}=\overline{c_1},\overline{c_2},...,\overline{c_N}$,其中,$\overline{C}$ 的第 i 个元素

可以用公式(11)表示:

$$\overline{c_i}=\frac{N}{n}\sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i}c_j \tag{11}$$

设一元时间序列 $Q=\langle q_1,q_2,\ldots,q_n\rangle$,$C=\langle c_1,c_2,\ldots,c_n\rangle$,通过公式(11)把其转化为 $\overline{Q},\overline{C}$,定义序列 $\overline{Q},\overline{C}$ 的距离为

$$DR(\overline{Q},\overline{C})=\frac{n}{N}\sum_{i=1}^{N}|\overline{q_i}-\overline{c_i}| \tag{12}$$

并且有公式(13)成立,证明过程见文献[20]:

$$DR(\overline{Q},\overline{C})\leqslant\sum_{i=1}^{n}|q_i-c_i| \tag{13}$$

3.2 索引结构的建立

对于长度为 n 的一元时间序列,如果直接用索引进行组织,由于维度过高,会使索引的性能严重退化[21,22].为此,可以使用 PAA 方法把序列从 n 维约减至 N 维($N<<n$),再用索引对 N 维向量进行组织.*LB_DTW* 的输入是 n 维原始序列,下面再提出一种下界距离 *LB_PAA*,其输入为 N 维向量(原始序列的 PAA 形式),这样便能够在 N 维索引结构上实现相似性搜索.

对一元查询序列 Q 的上、下边界序列 U 和 L,分别使用 PAA 方法表示为 $\overline{U},\overline{L}$:

$$\overline{u_i}=\frac{N}{n}\sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i}u_j \tag{14}$$

$$\overline{l_i}=\frac{N}{n}\sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i}l_j \tag{15}$$

下界距离 *LB_PAA* 定义为

$$LB\_PAA(\overline{U},\overline{L},\overline{C})=\sum_{i=1}^{N}\frac{n}{N}\begin{cases}|\overline{c_i}-\overline{u_i}|, & \text{if } \overline{c_i}>\overline{u_i}\\ |\overline{c_i}-\overline{l_i}|, & \text{if } \overline{c_i}<\overline{l_i}\\ 0, & \text{otherwise}\end{cases} \tag{16}$$

由公式(13)可知:

$$LB\_PAA(\overline{U},\overline{L},\overline{C})\leqslant LB\_DTW(U,L,C) \tag{17}$$

使用 R-Tree 对 N 维向量进行组织,设 V 为索引上的叶子结点,$R=(L,H)$表示与叶子结点 V 相关的 MBR,其中,$H=(h_1,h_2,\ldots,h_N)$,$L=(l_1,l_2,\ldots,l_N)$分别表示最小边界矩形 R 的上、下边界,R 中包含着满足上、下边界条件的 N 维向量.Q 与 R 的距离 Min*Dist*$(\overline{U},\overline{L},R)$ 定义为

$$\text{Min}Dist(\overline{U},\overline{L},R)=\sum_{i=1}^{N}\frac{n}{N}\begin{cases}|\overline{l_i}-\overline{u_i}|, & \text{if } \overline{l_i}>\overline{u_i}\\ |\overline{h_i}-\overline{l_i}|, & \text{if } \overline{h_i}<\overline{l_i}\\ 0, & \text{otherwise}\end{cases} \tag{18}$$

Min*Dist*$(\overline{U},\overline{L},R)$ 表示查询序列 Q 与 R 中所有 N 维向量的最小 *LB_PAA* 距离,证明过程见文献[13].因此,如果 Min*Dist*$(\overline{U},\overline{L},R)$ 大于阈值ε,则 Q 与 R 中所有 N 维向量的 *LB_PAA* 距离都大于ε,从而实现剪枝过滤功能.

LB_DTW,*LB_PAA* 和 Min*Dist* 均是针对等长一元时间序列的 DTW 下界距离,但应用场合不同,它们的输入对象分别为一元时间序列、N 维向量(序列的 PAA 形式)和 R-Tree 索引上的 MBR.三者之中,*LB_DTW* 是基础,*LB_PAA* 和 Min*Dist* 是对 *LB_DTW* 的延伸.

3.3 相似性搜索算法

下面给出ε范围搜索算法 *RangeSearch*(Q,ε,*rootNode*),算法以一元查询序列 Q、距离阈值ε和 R-Tree 的根结

点 *rootNode* 作为输入,采用结点递归的方式进行搜索,如图 7 所示.

算法. *RangeSearch*(Q,ε,P).
设U,L分别为Q的上、下边界序列,$\bar{U},\bar{L}$ 分别为其PAA形式
1. if P is a non-leaf node
2. for each child node T of P
3. if Min*Dist*($\bar{U},\bar{L},R$) //R是结点T对应的MBR
4. *RangeSearch*(Q,ε,T);
5. end
6. else
7. for each PAA point C in P
8. if $LB\_PAA(\bar{U},\bar{L},\bar{C})\leqslant\varepsilon$
9. Retrieve full sequence C from database;
10. if $DTW(Q,C)\leqslant\varepsilon$
11. Add C to Result;
12. end
13. end

Fig.7 Algorithm of ε range search

图 7 ε范围搜索算法

4 MTS 相似性搜索流程

上文在各个步骤上详细阐述了不等长 MTS 相似性搜索的实现方法,本节将从整体流程上对该方法进行描述.支持 DTW 距离的不等长 MTS 相似性搜索主要包括 3 方面的内容:

(1) 用索引结构组织 MTS 数据库,如图 8 所示;

(2) 提取查询序列的上、下边界特征,如图 9 所示;

(3) 用查询序列的上、下边界特征在索引结构上进行相似性搜索,如图 10 所示.

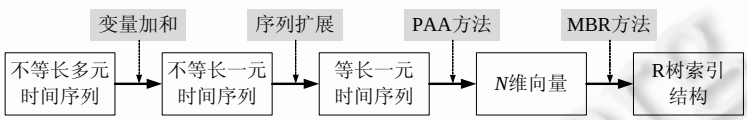


Fig.8 Process of index construction

图 8 索引构建流程

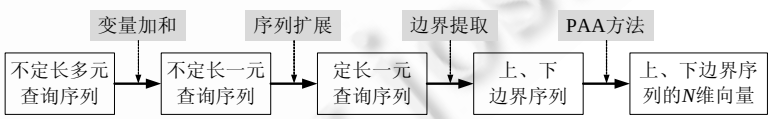


Fig.9 Feature extraction process of query upper and lower bounding series

图 9 查询序列上、下边界的特征提取流程

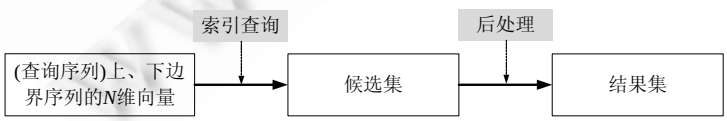


Fig.10 Process of similarity search

图 10 相似性搜索流程

使用变量加和方法,把数据库中的 MTS 转换为 UTS.但由于 UTS 的时间维数较高,直接用索引进行组织较为困难,因此在把 UTS 扩展为等长序列的基础上,通过 PAA 方法把等长 UTS 转换为 N 维向量,其中,N 为索引结

构能够处理的维度数.最后,使用 R-Tree 索引结构对所有的 N 维向量进行组织.

从 MTS 数据库中任意抽取一个序列作为查询序列,通过变量加和、序列扩展的方法,将其转换为定长一元时间序列,使得查询序列与 MTS 数据库中其他序列具有相同的长度;然后,提取定长一元时间序列的上、下边界序列;最后,通过 PAA 方法把上、下边界序列分别转换为 N 维向量,作为查询序列在索引查询中的特征.

在前两个流程的基础上,根据公式(16)、公式(18),使用查询序列上、下边界序列的 N 维向量形式,在索引结构上执行搜索,搜索结果构成候选集.

由于使用 DTW 下界距离在索引结构上执行搜索,候选集中不会产生漏报但会引入误报序列,因此必须对候选集进行后处理,即把候选集中的 UTS 映射为原始 MTS 后,依次计算每个原始 MTS 与多元查询序列的 DTW 距离,去除误报序列.由于候选集中的序列数量远小于数据库中的序列数量,因此能够提高搜索效率.

5 实验分析与讨论

5.1 实验环境与实验数据

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.选取两组多元时间序列数据集作为研究对象:Australian Sign Language[23]和 FlightData.

Australian Sign Language(记为 ASL)是一组手语信号数据集,包含 22 个连续型变量,左、右手的动作特征各用 11 个变量刻画:6 个变量(分别对应 6 个自由度)表示手所处的位置,5 个变量表示各手指的弯曲程度.手语信号数据集包含 95 种语意(95 个类),每种语意都有 27 组序列.不失一般性,选取前 8 种语意对应的序列作为实验数据集(记为 ASL),一共 216 个实验样本.8 种语意分别为 alive,all,answer,boy,building,buy,change-mind,cold,216 组多元时间序列的时间跨度在 47~95 之间,每组序列都体现一个完整的手语动作过程.

FlightData 是一组飞行数据集,它记录了某型飞机在训练过程中的飞行品质.为了便于实验分析,邀请飞参领域专家,通过专业软件截取飞行过程中表征特定飞行动作的数据段作为研究对象.5 组飞行动作分别为:加力盘旋、减速盘旋、水平横滚、180°盘旋和 360°盘旋,每组动作包含 200 个样本序列.飞行速度、飞行高度、俯仰角、横滚角和航向角这 5 个变量基本能够对这些飞行动作进行完整刻画.1 000 组多元飞行数据的时间跨度在 240~319 之间,每组序列都体现一个完整的飞行动作过程.

使用文献[24]中的共同主成分方法分别对两组数据集进行降维,方差贡献率参数 σ=80%,两组 MTS 数据集降维后的变量数均为 2,分别记为 ASL_DR,FlightData_DR,后续的实验针对降维后的数据进行讨论.

5.2 实验结果与分析

设多元时间序列 Q,C,使用变量加和方法转换为不等长一元时间序列 Q',C',序列扩展后形成等长一元时间序列 $Q'^{+},C'^{+},U'^{+},L'^{+}$是 Q'^{+}的上、下边界序列,$\overline{U'^{+}},\overline{L'^{+}}$ 和 $\overline{C'^{+}}$ 分别是 U'^{+},L'^{+}和 C'^{+}的 PAA 形式.

实验 1. MTS 转化为 UTS 时,变量加和方案的选择.

由性质 4 知,对于变量数为 $m(m>1)$的多元时间序列 Q,C,把任意对应的 $k(1\leqslant k\leqslant m)$个变量相加,能够得到两组一元时间序列 Q',C',每一个 k 值都对应着 C_m^k 种加和方案,因此,变量加和的方案一共有 $\sum_{k=1}^{m}C_m^k$ 种.下面研究把 MTS 转化为 UTS 时,如何选择最佳的变量加和方案.距离保持率 s 定义为

$$s=\frac{DTW(Q',C')}{DTW(Q,C)},s\subset[0,1] \tag{19}$$

它表示变量加和方法对 DTW 距离的保持程度,并用其评价各种方案的优劣.s 越大,说明转换效果越好.

设 MTS 数据集 *DsMts* 中含有 n 个序列,使用一种变量加和方案,把其转化为 UTS 数据集 *DsUts*.从 *DsUts* 中任意选择一个序列作为查询序列 Q',它在 *DsMts* 中对应的序列为 Q;*DsUts* 中的其他序列为 C_i' $(1\leqslant i\leqslant n-1)$,它在 *DsMts* 中对应的序列记为 C_i.s_i 表示 Q'与其他 $n-1$ 个序列 C_i' 的平均距离保持率,见公式(20):

$$s_i=\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{DTW(Q',C_i')}{DTW(Q,C_i)} \tag{20}$$

使用留一交叉验证法重复以上实验,可以得到 n 个平均距离保持率.则平均距离保持率的数学期望 s^*可由公式(21)确定,并用其作为变量加和方案的比较依据:

$$s^* = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{21}$$

下面针对降维数据集 FlightData_DR,ASL_DR 计算各种变量加和方案的 s^*.两种数据集中的变量数均为 2,不妨记为 x_1,x_2,则一共存在 3 种加和方案,结果见表 1.

Table 1 Comparison of different variable addition schemes

表 1 变量加和方案的比较

方案代号	变量加和方案	S^* (FlightData_DR)	S^* (ASL_DR)
方案 1	x_1	0.121 1	0.572 4
方案 2	x_2	0.788 7	0.379 6
方案 3	x_1+x_2	0.856 4	0.673 7

从实验结果可以看出,对于 FlightData_DR,ASL_DR,最优加和方式均为方案 3.使用最优加和方案得到的 UTS 数据集分别记为 FlightData_DR_UTS,ASL_DR_UTS.为了形象地验证性质 4,在 ASL_DR_UTS 中选择第 1 个序列作为查询序列 Q',分别计算 Q'与其他序列 C_i' (2≤i≤216)的距离 $DTW(Q',C_i')$;再求出 ASL_DR 中对应序列的距离 $DTW(Q,C_i)$,结果如图 11 所示.为了便于观察,图中仅截取了曲线的前 50 个点.可以看出,$DTW(Q,C_i)$的值都在 $DTW(Q',C_i')$ 之上,从而验证了性质 4 的正确性.

实验 2. 不等长 UTS 列扩展为等长 UTS.

使用序列扩展方法,把实验 1 中变量加和后的 UTS 数据集 FlightData_DR_UTS,ASL_DR_UTS 分别转化为等长 UTS 数据集,记为 FlightData_DR_UTS_ELen(其中的序列长度均为 320),ASL_DR_UTS_ELen(其中的序列长度均为 96).

下面验证序列扩展方法的有效性.

在 ASL_DR_UTS_ELen 中,选择第 1 个序列作为查询序列 Q'^{+},分别计算 Q'^{+}与数据集中其他序列 $C_i'^{+}$ (2≤i≤216)的距离 $DTW(Q'^{+},C_i'^{+})$;再求出 ASL_DR_UTS 中对应序列的距离 $DTW(Q',C_i')$,结果如图 12 所示.为了便于观察,图中仅截取了曲线的前 50 个点.

可以看出,$DTW(Q',C_i')$ 的值都在 $DTW(Q'^{+},C_i'^{+})$ 之上,从而验证了 $DTW(Q'^{+},C_i'^{+}) \leqslant DTW(Q',C_i')$.

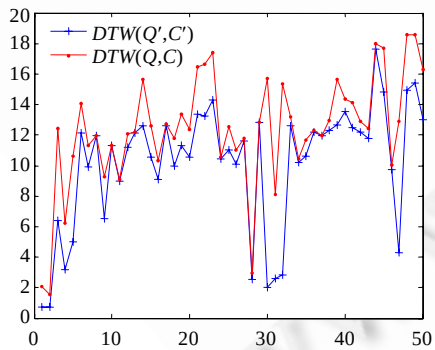


Fig.11 Validation of variable addition method

图 11 变量加和方法的有效性验证

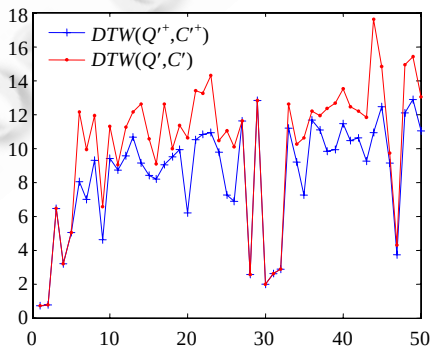


Fig.12 Validation of time series extended method

图 12 序列扩展方法的有效性验证

实验 3. 下界距离 *LB_DTW* 的紧致性分析.

紧致性越好,说明下界距离越接近实际距离,使用下界距离进行查询时,得到的误报序列就越少.实验用紧缩率和修剪率两个指标度量下界距离 *LB_DTW* 的紧致性.下界距离 *LB_DTW* 的紧缩率 T_{DTW} 定义为

$$T_{DTW}=\frac{LB\_DTW(U'^{+},L'^{+},C'^{+})}{DTW(Q'^{+},C'^{+})},T_{DTW}\subset[0,1] \tag{22}$$

修剪率 P 定义为

$$P=\frac{N_0}{N},P\subset[0,1] \tag{23}$$

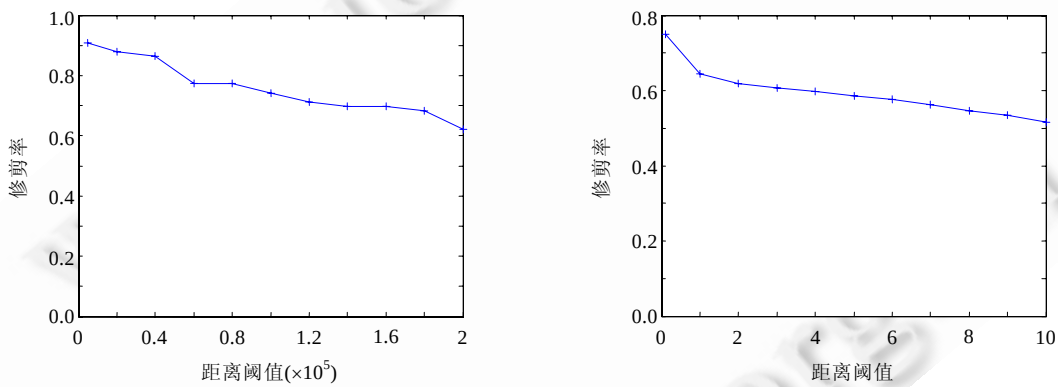
其中,N 表示使用顺序扫描方法与查询序列进行 DTW 距离计算的序列数,即为数据集中的序列数量;N_0 表示在下界距离 LB_DTW 的过滤作用下,不需要与查询序列进行 DTW 距离计算的序列数.

紧缩率、修剪率越高,表明紧致性越好,下界距离的过滤作用越明显,从而减少后处理的计算量,提高查询效率.下面以等长 UTS 数据集 FlightData_DR_UTS_Elen,ASL_DR_UTS_ELen 为实验对象,使用留一交叉验证法计算两组数据集中紧缩率 T_{DTW} 的平均值以及不同距离阈值 ε 下的修剪率,结果分别见表 2 并如图 13 所示.

Table 2 Tightness ratio of *LB_DTW*

表 2 下界距离 *LB_DTW* 的紧缩率

实验数据集	紧缩率的平均值(%)
FlightData_DR_UTS_ELen	66.32
ASL_DR_UTS_ELen	55.21



(a) FlightData_DR_UTS_ELen 中的实验结果 (b) ASL_DR_UTS_ELen 中的实验结果

Fig.13 Pruning efficiency of the lower bounding function *LB_DTW*

图 13 下界距离 *LB_DTW* 的修剪率

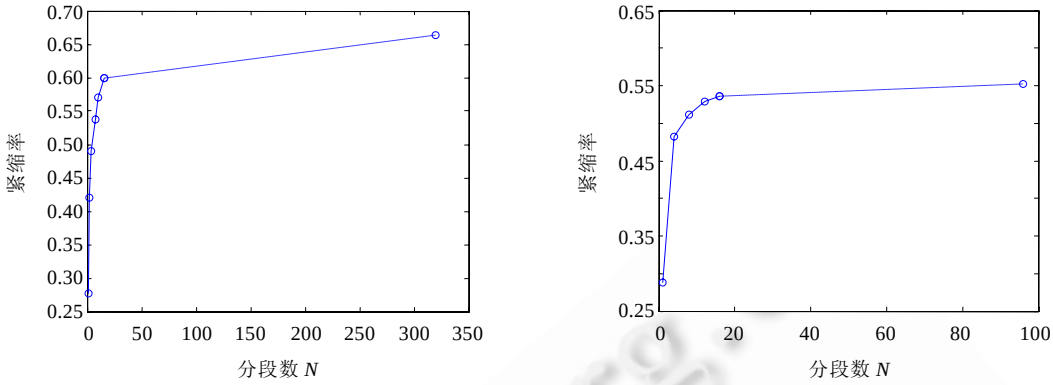
实验结果表明:针对两组数据集时,下界距离 LB_DTW 的紧缩率均大于 50%;距离阈值较小时,修剪率都能达到 70%以上.这说明下界距离 LB_DTW 的紧致性较好,在查询中的过滤作用较为明显.距离阈值越小,修剪率越高;随着阈值的不断增加,修剪率逐渐降低.这是因为下界距离 LB_DTW 的修剪率依赖于距离阈值,当距离阈值大于实际 DTW 距离时,下界距离将失去过滤作用.

实验 4. PAA 方法中分段数 N 的确定.

下面讨论把 UTS 表示为 PAA 形式时,分段数 N 的确定方法.下界距离 LB_PAA 的紧缩率 T_{PAA} 定义为

$$T_{PAA}=\frac{LB\_PAA(\overline{U'^{+}},\overline{L'^{+}},\overline{C'^{+}})}{DTW(Q'^{+},C'^{+})},T_{PAA}\subset[0,1] \tag{24}$$

分段数 N 决定着 $\overline{U'^{+}},\overline{L'^{+}},\overline{C'^{+}}$ 对 U'^{+},L'^{+},C'^{+} 的近似程度,N 越大,近似程度越高.以等长 UTS 数据集 FlightData_DR_UTS_Elen,ASL_DR_UTS_ELen 为实验对象,当 N 取不同值时,使用留一交叉验证法计算紧缩率 T_{PAA} 的平均值,结果如图 14 所示.



(a) FlightData_DR_UTS_ELen 中的实验结果 (b) ASL_DR_UTS_ELen 中的实验结果

Fig.14 Relation between *N* in PAA method and tightness ratio

图 14 PAA 方法中分段数 *N* 与紧缩率之间的关系

从实验结果可以看出,分段数 *N* 越大,紧缩率越高.这是由于随着 *N* 的增加,PAA 方法对序列的近似程度逐步提高,下界距离 *LB_PAA* 就越接近实际 DTW 距离.

针对 FlightData_DR_UTS_Elen,ASL_DR_UTS_ELen,*N* 分别取 320,96 时,下界距离 *LB_PAA* 的紧缩率与 *LB_DTW* 相同.这是因为当分段数 *N* 与时间序列的长度相同时,序列的 PAA 形式就是序列自身,*LB_DTW*, *LB_PAA* 具有相同的表达式.

文献[21,22]表明,对于高维空间索引结构,当维度数大于 16 时,索引的性能会严重下降.因此,一般取 $N\leqslant 16$. 从图 14 可看出,当 $N=16$ 时,紧缩率已经与最大值比较接近.

实验 5. 下界距离 *LB_DTW* 与 DTW 距离计算复杂度的比较.

通过理论分析可知,下界距离 *LB_DTW* 的计算复杂度低于 DTW 距离,下面用实验进行验证.

以 FlightData_DR_UTS_Elen,ASL_DR_UTS_ELen 为实验对象,用平均查询时间表示计算复杂度.为了消除实验环境引起的偏差,以两种方法计算时间的比值作为比较依据,结果见表 3.

Table 3 Comparison of computation time between *LB_DTW* and DTW distance

表 3 *LB_DTW*,DTW 距离计算时间的比较

实验数据集	t_{LB_DTW}/t_{DTW}
FlightData_DR_UTS_ELen	小于 1%
ASL_DR_UTS_ELen	小于 1%

从实验结果可以看出,针对两组数据集,下界距离 *LB_DTW* 的计算时间不足 DTW 距离的 1%.这是由于 *LB_DTW* 在形式和计算方法上都非常类似于 Minkowski 距离,序列间的点对匹配关系明确,计算过程中不需要考虑动态时间弯曲的影响.

6 结束语

本文给出了一种 MTS 的 DTW 下界距离 *LB_DTW*;然后以其为基础,提出了一种支持 DTW 距离度量的 MTS 索引结构,进而给出相应的相似性搜索算法.*LB_DTW* 以相关性质为理论支撑,实现了从一元向多元、从等长向不等长的拓展,把研究对象从等长一元时间序列推广到不等长多元时间序列,可视为对 *LB_Keogh* 方法的拓展.实验结果表明,本文提出的 MTS 索引方法能够有效地支持 DTW 距离的相似性搜索,且具有非漏报性,与顺序扫描方法相比,能够有效地提高搜索效率.

致谢 感谢 Mohammed Waleed Kadous 提供的实验数据集 Australian Sign Language.

References:

[1] Zhou DZ, Wu XL, Yan HC. An efficient similarity search for multivariate time series. Computer Applications, 2008,28(10): 2541−2543 (in Chinese with English abstract).

[2] Zhu Q, Wang XY, Keogh E, Lee SH. An efficient and effective similarity measure to enable data mining of petroglyphs. Data Mining and Knowledge Discovery, 2011,23(1):91−127. [doi: 10.10s07/s10618-010-0200-z]

[3] Li ZX, Zhang FM, Li KW. Research on pattern matching method for multivariate time series. Control and Decision, 2011,26(4): 565−570 (in Chinese with English abstract).

[4] Zhou DZ, Jiang WB, Li MQ. Efficient clustering algorithm for multivariate time series. Computer Engineering and Applications, 2010,46(1):137−139 (in Chinese with English abstract).

[5] Fu AWC, Keogh E, Lau LYH, Ratanamahatana CA, Wong RCW. Scaling and time warping in time series querying. The VLDB Journal, 2008,17:899−921. [doi: 10.1007/s00778-006-0040-z]

[6] Bankó Z, Abonyi J. Correlation based dynamic time warping of multivariate time series. Expert Systems with Applications, 2012, 39(2012):12814−12823. [doi: 10.1016/j.eswa.2012.05.012]

[7] Bhaduri K, Zhu Q, Oza NC, Srivastava AN. Fast and flexible multivariate time series subsequence search. In: Proc. of the 2010 IEEE Int'l Conf. on Data Mining. 2010. 48−57. [doi: 10.1109/ICDM.2010.36]

[8] Vlachos M, Hadjieleftheriou M, Gunopulos D, Keogh E. Indexing multidimensional time series. The VLDB Journal, 2006,15(1): 1−20. [doi: 10.1007/s00778-004-0144-2]

[9] Wong TSF, Wong MH. Efficient subsequence matching for sequences databases under time warping. In: Proc. of the 7th Int'l Database Engineering and Applications Symp. 2003. [doi: 10.1109/IDEAS.2003.1214921]

[10] Yi BK, Jagadish HV, Faloutsos C. Efficient retrieval of similar time sequences under time warping. In: Proc. of the 14th Int'l Conf. on Data Engineering. 1998. 201−208. [doi: 10.1109/ICDE.1998.655778]

[11] Kim SW, Sanghyun P, Chu WW. An index-based approach for similarity search supporting time warping in large sequence databases. In: Proc. of the 17th Int'l Conf. on Data Engineering. 2001. 607−614. [doi: 10.1109/ICDE.2001.914875]

[12] Keogh E, Ratanamahatana CA. Exact indexing of dynamic time warping. Knowledge and Information Systems, 2005,7:358−386. [doi: 10.1007/s10115-004-0154-9]

[13] Zhu YY, Shasha D. Warping indexes with envelope transforms for query by humming. In: Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of Data. 2003. 181−192. [doi: 10.1145/872757.872780]

[14] Mu B, Yan JL. Efficient time series lower bounding technique. Computer Engineering and Applications, 2009,45(11):168−171 (in Chinese with English abstract).

[15] Berndt DJ, Clifford J. Using dynamic time warping to find patterns in time series. In: Proc. of the Workshop on Knowledge Discovery in Databases. 1994. 229−248.

[16] Zhou M, Wong MH. Efficient online subsequence searching in data streams under dynamic time warping distance. In: Proc. of the 24th Int'l Conf. on Data Engineering. 2008. 686−695. [doi: 10.1109/ICDE.2008.4497477]

[17] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases. In: Proc. of the ACM SICMOD Conf. on Management of Data. 1994. 419−429. [doi: 10.1145/191839.191925]

[18] Zhang MB, Lu F, Shen PW, Cheng CX. The evolvement and progress of R-tree family. Chinese Journal of Computers, 2005,28(3): 289−300 (in Chinese with English abstract).

[19] Guttman A. R-trees: A dynamic index structure for spatial searching. In: Proc. of the ACM SIGMOD. 1984. 47−57. [doi: 10.1145/602259.602266]

[20] Yi BK, Faloutsos C. Fast time sequence indexing for arbitrary *Lp* norms. In: Proc. of the 26th Int'l Conf. on Very Large Databases. 2000. 385−394.

[21] Park S, Chu WW, Yoon J, Won J. Similarity search of time-warped subsequences via a suffix tree. Information Systems, 2003,28: 867−883. [doi: 10.1016/S0306-4379(02)00102-3]

[22] Seidl T, Kriegel H. Optimal multi-step k-nearest neighbor search. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. 1998. 154−165.

[23] Kadous MW. High-Quality recordings of Australian sign language signs. 2002. http://kdd.ics.uci.edu/databases/High-quality Australian Sign Language/High-quality Australian Sign Language.html

[24] Yoon H, Yang K, Shahabi C. Feature subset selection and feature ranking for multivariate time series. IEEE Trans. on Knowledge and Data Engineering, 2005,17(9):1186−1198. [doi: 10.1109/TKDE.2005.144]

附中文参考文献:

[1] 周大镯,吴晓丽,闫红灿.一种高效的多变量时间序列相似查询算法.计算机应用,2008,28(10):2541−2543.

[3] 李正欣,张凤鸣,李克武.多元时间序列模式匹配方法研究.控制与决策,2011,26(4):565−570.

[4] 周大镯,姜文波,李敏强.一个高效的多变量时间序列聚类算法.计算机工程与应用,2010,46(1):137−139.

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

[18] 张明波,陆锋,申排伟,程昌秀.R 树家族的演变和发展.计算机学报,2005,28(3):289−300.

李正欣(1982－),男,河南信阳人,博士,讲师,主要研究领域为信息系统工程与智能决策,数据挖掘.
E-mail: lizhengxin_2005@163.com

李克武(1968－),男,副教授,主要研究领域为智能信息处理与决策.
E-mail: 314lkw@126.com

张凤鸣(1963－),男,教授,博士生导师,主要研究领域为信息系统工程与智能决策.
E-mail: zhangfm_2010@163.com

张晓丰(1978－),男,博士,副教授,主要研究领域为智能信息处理与决策.
E-mail: zhxfzhxf1@sina.com.cn