DOI:10.11992/tis.201507064
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160315.1239.012.html

分段聚合近似和数值导数的动态时间弯曲方法

李海林,梁叶
(华侨大学 信息管理系,福建 泉州 362021)

摘 要:针对动态弯曲方法对时间序列数据相似性度量的质量和效率的局限性,本文提出一种基于分段聚合近似和数值导数的动态时间弯曲方法。该方法通过分段聚合近似将时间序列数据进行有效地降维,再结合数值导数对降维后的特征序列构建新特征序列,并且设计符合该特征序列相似性度量方法。实验结果分析表明,与传统动态弯曲方法相比,新方法具有较好的度量质量,能在时间序列数据挖掘中得到较好的分类效果,且在低维空间具有较高的分类效率,具有一定的优越性。

关键词:动态时间弯曲;时间序列;分段聚合近似;数值导数;相似性度量;分类;数据降维;特征表示
中图分类号: TP301 **文献标志码**:A **文章编号**:1673-4785(2016)02-0249-08

中文引用格式:李海林,梁叶. 分段聚合近似和数值导数的动态时间弯曲方法[J]. 智能系统学报, 2016, 11(2): 249-256.
英文引用格式:**LI Hailin, LIANG Ye. Dynamic time warping based on piecewise aggregate approximation and data derivatives[J]. CAAI transactions on intelligent systems**, 2016, 11(2): 249-256.

Dynamic time warping based on piecewise aggregate approximation and data derivatives

LI Hailin, LIANG Ye
(Department of Information Management, Huaqiao University, Quanzhou 362021, China)

Abstract: Dynamic time warping (DTW) is often used to measure the similarity of time series data; however, it has efficiency and quality limitations. In this study, a novel DTW method combining piecewise aggregate approximation (PAA) and derivatives is proposed to measure the similarity of time series. The dimensionality of the time series data was effectively reduced by PAA, and the feature sequence was transformed into new sequences by combining the numerical derivatives after the dimensionality reduction. Furthermore, the DTW design corresponded to the similarity measurement method of the feature sequence. The experimental results demonstrate that the proposed method is superior because it has better measurement quality, obtains a better classification effect in time series data mining, and has high efficiency in lower dimensional spaces.

Keywords: dynamic time warping; time series; piecewise aggregate approximation; numerical derivative; similarity measure; classification; dimensionality reduction; feature representation

近年来,时间序列成为了数据挖掘研究领域中的热点,其研究成果已被广泛地应用在工业、金融、气象、航天、生物医学等各种领域[1-5]。然而,相似性度量方法成为时间序列数据挖掘领域中基础而又极其关键的工作之一,很多挖掘任务的结果易受其相似性度量质量和效率的影响。目前,动态时间弯曲(dynamic time warping, DTW)是一种鲁棒性较强的方法之一,其最早应用于语音识别[6],如今已被广泛应用于时间序列相似性的度量[7-9]。欧氏距离虽能快速度量时间序列之间的相似性,但其限于相同时间点上的数据匹配,而且对异常数据点较为敏感。DTW 不仅可以弯曲度量不等长时间序列之间

收稿日期:2015-07-24. **网络出版日期**:2016-03-15.
基金项目:国家自然科学基金项目(61300139);福建省中青年教育科研基金项目(JAS14024);华侨大学中青年教师科研提升计划项目(ZQN-PY220).
通信作者:李海林. E-mail:hailin@hqu.edu.cn.

的相似性,其度量质量能较好地反映数据之间的异步关系[10]。然而,由于 DTW 需要在代价矩阵中计算最优弯曲路径,使其具有较高的时间复杂度或较好的度量精度[11-13]。

时间序列降维表示是通过某种方法将原始序列进行转换或者特征提取,以达到将原始时间序列在低维度表示的目的。由于直接使用和计算原始时间序列可能需要付出较大的代价,如消耗较多存储空间,运行效率低等问题,加上原始时间序列包含很多“噪音点”,容易导致相似性度量结果不准确。因此,为简化数据模型和算法的复杂性,提高数据挖掘技术的性能,有必要对原始时间序列进行降维处理,达到效率和准确性的平衡。目前,针对时间序列降维的方法已有很多且较为成熟,如分段聚合近似(piecewise aggregate approximation, PAA)[14],该方法通过将时间序列进行平均分段,并以分段均值反映分段信息,最终达到数据降维的目的;分段线性近似(piecewise liner approximation, PLA)[15]利用线性模型对时间序列进行分割,根据不同的分割策略可以得到不同的时间序列降维效果;基于域变换的离散傅里叶变换(discrete fourier transform, DFT)[16]和离散小波变换(discrete wavelet transform, DWT)[17],利用变换后的系数对时间序列进行特征表示;奇异值分解(singular value decomposition, SVD)[18],利用数值计算将高维数据转换到低维空间,不仅应用在时间序列数据降维及索引,而且也被广泛地应用于模式识别、图像压缩等等。由于降维方法对整个研究过程起着十分关键的作用,不仅要求算法能够尽可能简单快速,以降低时间消耗,也要尽量充分反映时间序列的信息。因此研究过程中应选择合适的降维方法来解决问题。

鉴于传统 DTW 方法在度量时间序列相似性时具有计算时间效率较低和缺少考虑序列的形态(即凹凸性)等问题,并且从使用较为简单的数据降维方法以简化模型的角度出发,本文提出通过利用分段聚合近似(PAA)方法对时间序列进行数据降维,获得保持原始序列主要特征的数据序列,再构造基于数据点的一阶导数的新特征序列,结合动态时间弯曲提出一种新的距离度量方法,即近似导数动态时间弯曲方法。数值实验分类结果和效率分析表明,该方法能从形态视角出发,较好地从数据集中找出相似序列,提高数据挖掘领域中分类算法的质量,具有一定的优越性。

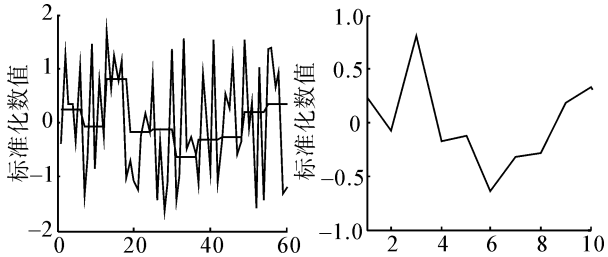
1 相关理论基础

1.1 分段聚合近似(PAA)

分段聚合近似是一种有效的数据降维方法,其在减小数据存储和提高计算效率方面起到了很大的作用。通过对长度为 n 的序列 $S=(s_1,s_2,\cdots,s_n)$ 转化为另一条长度为 m 的序列 $Q=(q_1,q_2,\cdots,q_m)$,实现时间序列的数据降维和特征表示,其中 $n>m$,且令 $k=n/m$。新序列中任意元素 q_i 满足:

$$q_i=\frac{1}{k}\sum_{j=k*(i-1)+1}^{k*i}s_i,\qquad 1\leqslant i\leqslant m \tag{1}$$

对序列 S 进行分段后求出每段的均值,每段均值形成序列 Q,进而达到序列 S 数据降维的目的。如图 1,子图(a)显示了长度为 60 的时间序列 S 通过 PAA 被平均分成 10 段,每段均值代表相应片段的特征;子图(b)显示由 10 个均值组成的新序列 Q。



(a)时间序列及其分段均值 (b)时间序列的特征表示

图 1 基于 PAA 的时间序列数据降维和特征表示

Fig.1 Dimensionality reduction and feature representation of time series based on PAA

分段聚合近似方法通过均值来表示序列片段的特征,容易忽略数据的局部形态变化情况。然而,在实际运用中,时间序列的整体形态通常是关注和研究的重点。PAA 得到的序列特征不仅能够很好地反映较长时间序列数据形态的整体变化趋势,而且还能对时间序列数据进行数据降维,起到提高相关数据挖掘算法效率的作用。

1.2 动态时间弯曲

动态时间弯曲最初被应用于语音识别中,常被运用到比较 2 条时间序列的相似性。针对 2 条时间序列 $S=(s_1,s_2,\cdots,s_n)$ 和 $Q=(q_1,q_2,\cdots,q_m)$,对任意两点之间的距离构建一个 $n\times m$ 的距离矩阵 $\boldsymbol{D}$,其中 $d(i,j)$ 表示时间序列数据点 s_i 和 q_j 之间的距离,即 $d(i,j)=(s_i-q_j)^2$。DTW 的基本思想就是从距离矩阵中寻找一条使得 2 条序列之间的累计距离最小的路径,其最小累积距离值为

$$\mathrm{DTW}(S,Q)=\min_{W}\left(\sum_{k=1}^{K}w_k\right) \tag{2}$$

弯曲路径是一条具有连续 K 个距离矩阵元素的集合 $W=(w_1,w_2,\cdots,w_K)$,第 k 个元素为 $w_k=(i,j)_k$ 且 $\max(n,m)\leqslant K\leqslant n+m-1$。与此同时,弯曲路径通常要遵循着 3 个条件:

1)边界性:$w_1=(1,1)$,$w_k=(n,m)$;

2)连续性:$w_k = (a,b)$, $w_K - 1 = (a',b')$,则 $a - a' \leqslant 1$, $b - b' \leqslant 1$;

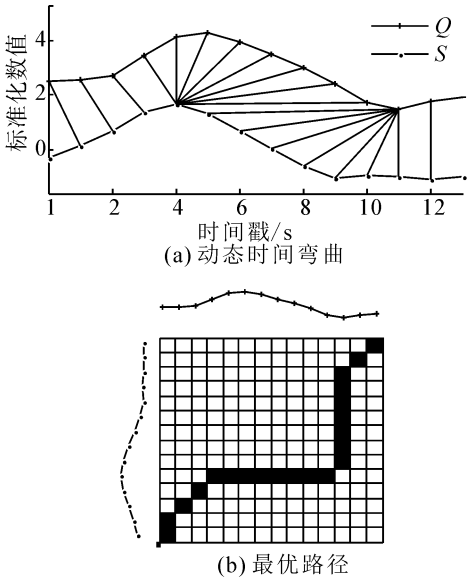
3)单调性:$w_k = (a,b)$, $w_K - 1 = (a',b')$,则 $a - a' \geqslant 0$, $b - b' \geqslant 0$;

最优弯曲路径的求解可以通过动态规划来找出最小累积距离,即

$$\gamma(i,j) = d(i,j) + \min\begin{cases}\gamma(i-1,j-1)\\ \gamma(i-1,j)\\ \gamma(i,j-1)\end{cases} \tag{3}$$

式(3)说明当前元素的累积距离 (i,j) 是当前元素距离值与邻近 3 个元素累积距离的最小值之和,且 $\gamma(n,m)$ 表示 DTW 度量时间序列 S 和 Q 的最小距离 $\mathrm{DTW}(S,Q) = \gamma(n,m)$ 。

如图 2(a)所示,动态时间弯曲方法可以很好地匹配不同时间点但具有相似形态的序列特征,使得时间序列的相似性特征能够较好地得到体现。另外,通过累积矩阵的构建可以获取其最优弯曲路径和最小度量距离,如图 2(b)所示。



(a)动态时间弯曲

(b)最优路径

图 2　动态时间弯曲及其最优路径

Fig.2　Dynamic time warping and the best path

DTW 在很多领域的应用中都有着很好的效果,然而较高的时间复杂度成为了该方法的瓶颈问题。目前也有部分学者对其进行了改进,使其具有较好的度量效率。另外,由于 DTW 有时也存在过度“变态”弯曲的状况,即一条时间序列的一个值对应着另一条序列的较长子序列,进而造成计算结果不精确的问题。针对该问题,Keogh 和 Pazzani 提出了导数动态时间弯曲(DDTW),有效地克服了 DTW 对时间轴的过度弯曲时造成的影响。

DDTW 与 DTW 两种方法基本算法相似,DTW 中距离矩阵元素 (i,j) 的距离值为 $d(i,j) = (s_i - q_j)^2$;DDTW 距离矩阵元素的距离值为对应导数之间的距离,即 $d(i,j) = (s_i' - q_j')^2$,且数值导数 s'_i 为

$$s'_i = \frac{(s_i - s_{i-1}) + (s_{i+1} - s_{i-1})/2}{2} \tag{4}$$

2　近似导数动态时间弯曲方法

若直接使用动态时间弯曲方法对时间序列进行相似性度量,则时间消耗太大,并且也会造成时间过度弯曲等问题,不利于大规模的长度较长的时间序列数据挖掘。为此,在使用动态时间弯曲度量相似性之前,通常先对时间序列数据进行数据降维,不仅可以得到反映时间序列大部分信息的低维特征表示,而且还能近似表示原始时间序列的整体性信息。除此之外,在分类过程中,针对低频时间序列数据缺少具体局部信息的状况,序列被归为哪一类通常可能取决于其整体形状,而不是局限于严格的数值比较。因此,相似性度量的结果不仅需要充分考虑序列的具体数值,也要考虑序列整体形态(凹凸性)。

传统研究方法是对时间序列进行数据降维,然后使用传统距离公式直接进行相似性度量。由于数据降维会损耗时间序列的部分信息,实验效果也就相当于以准确性来换取时间效率。鉴于时间序列具体数据值的重要性、形态的特殊性、度量准确性以及高运行效率的必要性,结合分段聚合近似以及数值导数,提出一种基于分段聚合近似和数值导数的动态时间弯曲方法,称为分段聚合导数距离(piecewise approximation derivative distance, PADD)。该方法主要综合考虑时间序列的具体数值信息和形态特征,以获得更好的度量效果,提高运行效率,减少异常点对度量的影响为目标,同时从时间序列的整体数据匹配和形态的凸凹性出发,更为完善地度量时间序列异步相关性。

时间序列 $S = (s_1, s_2, \cdots, s_n)$ 和 $Q = (q_1, q_2, \cdots, q_m)$,通过分段聚合近似方法对它们进行数据降维并进行整体数据的特征表示,即 $\bar{S} = \mathrm{PAA}(S, w_s)$ 和 $\bar{Q} = \mathrm{PAA}(Q, w_q)$,其中 w_s 和 w_q 为降维后维度,也是分段数目。另外,为了反映原始时间序列整体形态的特征,利用数据导数来反映 PAA 对原始时间序列降维后的序列进行整体形态特征描述。利用式(4)对 $\bar{S}$ 和 $\bar{Q}$ 进行数据导数计算,以反映原始时间序列的整体形态特征,即 S' 和 Q' 。最后,利用动态时间弯曲方法分别对($\bar{S}$, $\bar{Q}$)和(S' , Q')进行距离度量,并综合两者的加权平均作为新方法的距离度量

函数,即

$$D_{\mathrm{PADD}}(S,Q)=\alpha\mathrm{DTW}(\bar{S},\bar{Q})+\beta\mathrm{DTW}(S',Q') \tag{5}$$

式中:α、β 是通过

$$\begin{cases}\alpha=\cos^2\theta\\ \beta=\sin^2\theta\end{cases},\quad \theta\subset[0,\frac{\pi}{2}] \tag{6}$$

求解得到。如图 3 所示,参数 α、β 可以通过三角公式求解得到,两者取值范围均在[0,1],且式(6)确保了两者之和为 1,使其特征序列的动态时间弯曲距离是线性组合关系。为方便起见,系数 α、β 可由参数 θ 来确定。

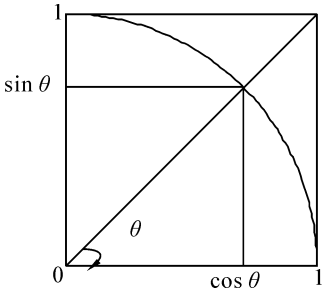


图 3 参数 α、β 选取原理

Fig.3 The principle of selecting parameters α and β

结合上述思想,本文提出的分段聚合导数距离(PADD)算法步骤如下:

输入 时间序列 $S=(s_1,s_2,\cdots,s_n)$ 和 $Q=(q_1,q_2,\cdots,q_m)$,分段数目 w_s 和 w_q 以及参数 θ;

输出 度量距离 $D_{\mathrm{PADD}}(S,Q)$。

1)分别对时间序列进行平均分段,每个子序列长度分别为 $k_s=n/w_s$ 和 $k_q=m/w_q$。

2)根据 PAA 思想,对每个子序列进行均值表示,获得 PAA 序列 $\bar{S}=(\bar{s}_1,\bar{s}_2,\cdots,\bar{s}_{w_s})$ 和 $\bar{Q}=(\bar{q}_1,\bar{q}_2,\cdots,\bar{q}_{w_q})$,其中两组特征序列的元素分别为 $\bar{s}_i=\frac{1}{k_s}\sum_{j=k_s(i-1)+1}^{ik_s}s_j$ 和 $\bar{q}_i=\frac{1}{k_q}\sum_{j=k_s(i-1)+1}^{ik_q}q_j$。

3)利用式(4)分别对 PAA 序列进行数据导数求解,得到基于导数的特征序列 $S'=(s'_1,s'_2,\cdots,s'_{w_s-2})$ 和 $Q'=(q'_1,q'_2,\cdots,q'_{w_q-2})$。

4)利用 DTW 分别对上述两种特征序列进行度量距离,并实现线性组合,即 $D_{\mathrm{PADD}}(S,Q)=\alpha\mathrm{DTW}(\bar{S},\bar{Q})+\beta\mathrm{DTW}(\bar{S}',\bar{Q}')$。

在上述算法中,1)实现了时间序列的分段,使用等划分的方法实现数据划分并获得相应的子序列片段;2)中方法对子序列段的均值特征表示,而且大量研究结果表明[19],对于高维时间序列来说,PAA 方法可以保留足够的数据信息来反映原序列的波动形态;3)实现均值特征序列的导数表示,以便更好地反映特征序列的形态变化情况。最后,利用式(6)同时从均值特征序列和导数特征序列来反映原始时间序列的数据信息,以确保距离度量函数能够较好地通过特征序列之间的数值关系来反映原始时间序列之间的数值和形态变化关系。

另外,从算法的整个描述过程,容易分析得到新方法的计算时间效率。长度为 n 的时间序列进行分段聚合近似时,其时间复杂度为 $O(n)$,计算一阶导数的时间复杂度也为 $O(n)$。另外,对 PAA 特征序列和导数序列进行 DTW 距离度量所需要的时间复杂度分别为 $O(w_sw_q)$ 和 $O((w_s-2)(w_q-2))$,故 PADD 整个算法过程的时间复杂度近似为 $O(2w_sw_q+4n)$。由于分段数目 w_s 和 w_q 通常远小于原始时间序列的长度,即 $w_s\ll n$ 和 $w_q\ll n$,故 PADD 的时间复杂度 $O(2w_sw_q+4n)$ 要低于与传统 DTW 和 DDTW 的时间复杂度 $O(n^2)$。特别地,数值实验的分类结果和时间计算效率分析表明,PADD 能够较好地度量时间序列的相似性,反映数据形态之间的变化关系,进而提高时间序列数据挖掘算法的分类结果,还能在低维特征空间下获得较高的时间效率。

3 数值实验

为了更好地理解和检验 PADD 方法的性能,实验通过对时间序列数据集进行分类实验,与传统 DTW 和 DDTW 方法进行比较,验证新方法在不同时间序列数据集中的距离度量质量和时间计算效率。

3.1 分类实验

实验一共进行两个阶段,1)训练模型,使用“留一法”(leave-one-out)求出合适的参数 θ 构建度量距离;2)测试模型,得到分类效果。由于时间序列取值范围差距较大,度量相似性结果可能会出现很大偏差,因此有必要对所有时间序列数据进行归一化处理,消除量纲。

使用训练集训练模型过程中,给定时间序列分段数目 w 以及范围为 $[0,\pi/2]$ 且步长为 0.01 的 θ,利用分段聚合近似以及一阶求导分别得到新的特征序列。训练阶段结合“枚举法”测试在 w 情况下每个 θ 的分类情况,选择错误率最小的 θ 作为该 w 情况下的最优参数。一旦求得参数立即构建度量距离公式(5),进入 2)利用最近邻(1-NN)分类方法检测该距离公式分类效果。

图 4 给出了数据集 ECG 在不同 w 取值下的最佳参数 θ。根据式(5)与式(6)易知,θ 取值越大,

时间序列相似性度量形态匹配所占比重越大;反之,则形态匹配的比重就越小。图 4 也说明了时间序列数据集在大部分情况下更倾向于形态匹配而不是具体数值上的比较。

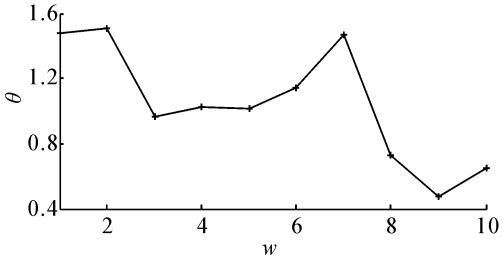


图 4　不同 w 下的最优参数 θ

Fig.4　The best parameter θ for different w

本次实验使用 Keogh 教授提供的数据集[20],并随机抽取 20 个数据集进行分类实验。如表 1 所示,包含了各个数据集的基本详细信息,包括数据集名称、类别个数、训练集个数、测试集个数、时间序列长度。表 2 包含了 DTW、DDTW、PADD 的平均错误率,并且给出了 PADD 的最小错误率,其中括号表示能够取得 PADD 的最小分类错误率的 w,例如当 CBF 数据被降维到原始序列长度的 70%时,可以取得最小错误率 0。

表 1　时间序列数据集信息

Table 1　Information of time series dataset

Datasets	NO.C	S.tr	S.Tst	length
Beef	5	30	30	470
CBF	4	30	900	128
Coffee	2	28	28	286
D.S.Z.	4	16	306	345
ECG200	2	100	100	96
E.F.D.	2	23	861	136
Face Four	4	24	88	350
Fish	7	175	175	463
Gun Point	2	50	150	150
I.P.D.	2	67	1 029	24
Lighting7	7	70	73	319
M. I.	10	381	760	99
M.S.	2	20	1 252	84
Olive Oil	4	30	30	570
OSU Leaf	6	200	242	427
S.A.R.S.	2	20	601	70
S.A.R.S II	2	27	953	65
Symbols	6	25	995	398
S.C.	6	300	300	60
Trace	4	23	1 139	82

注:表头 NO.C、S.Tr、S.Tst、Length 分别表示类别个数、训练集个数、测试集个数、序列长度。数据集名 D.S.Z.、E.F.D.、I.P.D.、M.I.、M.S.、S.A.R.S.、S.A.R.S II、S.C.分别表示 Diatom Size Reduction、ECG Five Days、Italy Power Demand、Medical Images、Mote Strain、Sony AIBO Robot Surface、Sony AIBO Robot Surface II、Synthetic Control.

表 2　实验分类错误率

Table 2　Classification error ratesof the experiments　%

Datasets	DTW	DDTW	PADD(avg)	PADD(min)
Beef	50.00	30.00	32.33	23.33(2)
CBF	0.33	27.11	1.14	0.00(7)
Coffee	17.86	14.29	2.86	0.00(3,5,10)
D.S.Z.	3.27	6.21	5.49	3.27(5,10)
ECG200	23.00	19.00	12.90	9.00(2)
E.F.D.	23.23	35.08	26.82	17.65(1)
Face Four	17.05	29.55	20.91	15.91(5,6,7)
fish	16.57	10.29	12.41	10.86(6)
Gun-Point	9.33	1.33	5.40	2.67(5,7)
I.P.D.	4.96	8.75	6.81	4.28(6,8)
Lighting7	27.40	41.09	28.63	23.29(1)
M.I.	26.32	32.24	29.89	26.45(5)
M.S.	16.53	28.04	16.53	12.70(8)
Olive Oil	13.33	16.67	14.00	13.33(2~9)
OSU Leaf	40.91	7.02	27.89	19.42(2,5)
S.A.R.S.	27.45	25.46	28.40	20.13(9)
S.A.R.S II	16.98	16.37	21.49	16.26(9)
Symbols	5.03	2.51	4.31	3.62(2,3)
S.C.	0.67	56.33	6.33	0.67(10)
Trace	0.00	1.00	2.90	0.00(5,10)
MEAN	17.01	20.42	15.37	11.14
SD	13.42	14.74	10.78	8.89

从实验结果来看,PADD 的平均分类效果与 DTW 和 DDTW 相比有着较大的优势,并且从最小分类错误率的均值及标准差上来看也取得了不错的效果,进而验证了 PADD 在时间序列数据挖掘中距离度量的有效性。实验结果反映了时间序列分段聚合后的特征序列长度大多数为原始序列长度的 50%~100%,即 w 大多数取值为 5~10 时,PADD 将会取到比较好的效果。这也说明在进行序列转换过程中,将时间序列的整体信息以及具体细节尽量完整保存下来会取得更好的效果。

为了更为清楚地剖析分类结果在不同参数下的情况,给出了 Sony AIBO Robot Surface 和 ECG 这两个数据集的分类情况,如图 5 所示。从图 5 中可以看出,分类错误率呈高低波动状态。对于某些数据集如 Sony AIBO Robot Surface 来说,在不同 w 情况下并非所有的 PADD 错误率都是最小的,但是在最

小错误率上却能够优于 DTW 和 DDTW 的错误率。然而,在有些数据集如 ECG,不管 w 取何值,新方法的分类错误率都是低于另外两者的错误率。

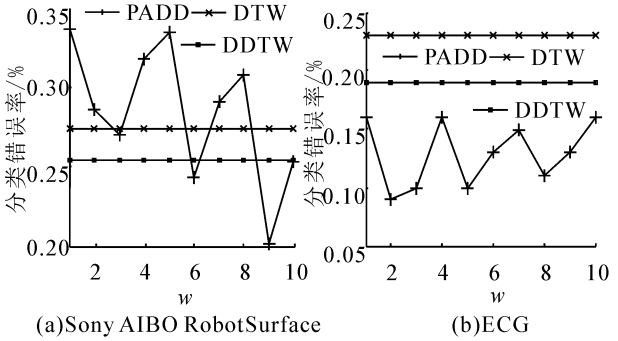


图 5 分类错误率和时间序列分段聚合长度的关系

Fig.5 The relationship of classification and the length of time series piecewise aggregation

从图 5 中发现,不同数据集呈现出来的性能不一样。由于分段聚合近似采用的是均值作为替代值,这并非最佳的降维方法,因此,若数据集中的序列数据波动较多、振幅较大,会对降维效果造成一定的影响。如图 6 所示。

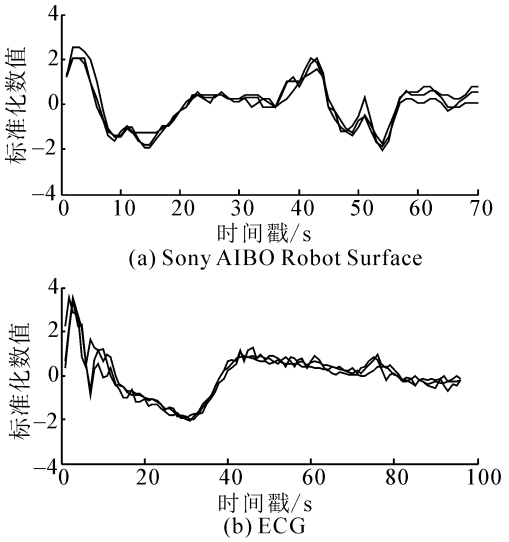


图 6 实验数据集序列示例

Fig.6 The example of datasets

图 6 分别给出的是训练集 Sony AIBO Robot Surface 和 ECG 的 3 条时间序列,可以观察到子图(a)的序列在整体波动上比子图(b)的多。因此,影响了序列降维的效果,给后续的相似性度量造成了一定的影响。由于数据集的特性会给本方法造成一定的影响,因此,本方法适合波动较为平缓的数据集,并会取得较好的效果。

图 7 描述的是 PADD 的平均分类错误率以及最小分类错误率与 DTW、DDTW 的错误率的比较,为了便于直观比较,数值均经过归一化后取值范围为[0,0.5],数值偏向方表示对应方法的错误率较大。

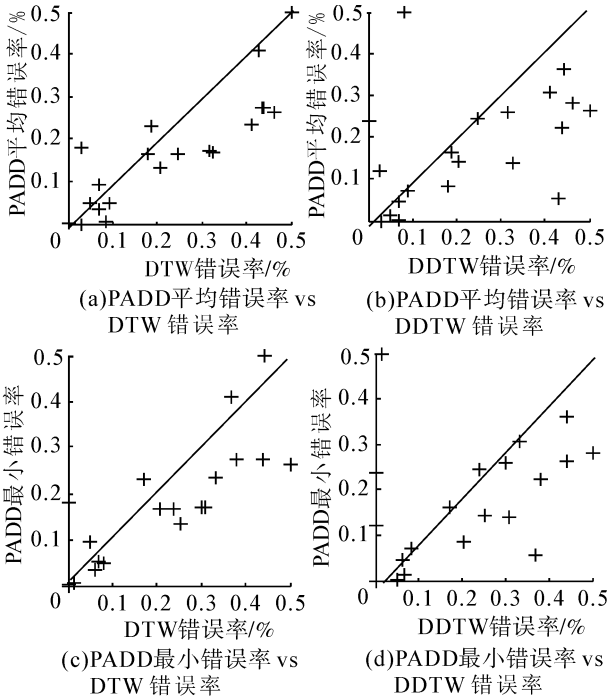


图 7 分类错误率的比较

Fig.7 Comparison of classification error rates

图 7 中子图(a)、(b)分别描述 PADD 的平均错误率与 DTW、DDTW 的错误率比较,子图(c)、(d)分别描述 PADD 最小错误率与 DTW、DDTW 错误率比较。结果分析表明,不管从平均错误率还是最小错误率角度来比较,PADD 错误率所对应的坐标纵轴值相对较小,使其偏向于 DTW 和 DDTW 所代表的横轴值较大,大多散点都偏向于 DTW 和 DDTW,故说明 PADD 具有较小的分类错误率,进行验证了本文新方法在时间序列度量中的有效性和优越性。

3.2 时间效率分析

本实验使用“留一法”(leave-one-out)求解参数,参数确定后立即构建度量距离公式。特征序列长度越长,计算所要消耗的时间越大,且成数量级增长。如图 8 所示,描述了数据集 Sony AIBO Robot Surface 和 ECG 在不同 w 条件下 PADD 时间效率与 DTW、DDTW 的时间效率比较。

由于 DDTW 需要利用式(4)预先对时间序列数据进行求导,因此 DDTW 时间效率略高于 DTW 时间效率。另外,随着特征序列长度 w 的增长,PADD 消耗的时间也随之增长。当分段数目 w 增长到一定程度时,PADD 的计算时间消耗会大于传统 DTW 和 DDTW 方法。理论上结合时间复杂度的分析可知,PADD 的时间复杂度近似为 $O(2w^2+4n)$,DTW 和 DDTW 的时间复杂度为 $O(n^2)$,当且仅当 $w<\sqrt{\dfrac{(n-2)^2-4}{2}}$ 时,PADD 的时间效率要低于 DTW

和 DDTW。因此,在分段比例超过约 $\sqrt{\frac{(n-2)^2-4}{2}}/n=\sqrt{\frac{n-2}{n}}$ 时,时间效率便开始降低。可以发现,n 为任意值,该比例均约为 70%,从图中也清楚地看到,DTW、DDTW、PADD 三者在分段比例约为 70%的地方相交。另外,在 PADD 过程中会涉及到一些辅助计算,故若大于某一特定降维之后的维度时,PADD 将会消耗更多的计算时间。然而,如图 8 所示,在数据降维后的低维空间中,PADD 可以取得较好的时间效率,具有一定的计算性能优势。

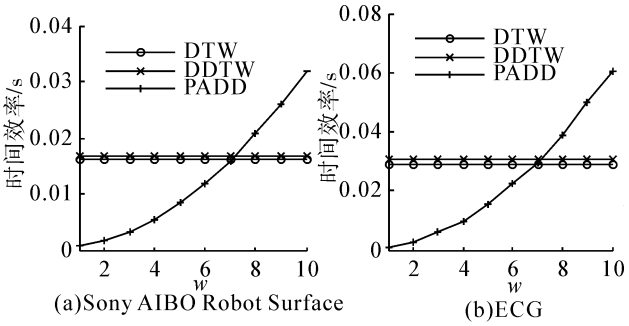


图 8　3 种方法的时间效率

Fig.8　Time efficiency of the three methods

4　结束语

针对动态弯曲方法对时间序列数据相似性度量质量和效率的局限性,以及从简化模型的角度考虑,本文提出一种基于分段聚合近似和数值导数的动态时间弯曲方法,结合分段聚合近似算法以及数值求导将时间序列进行转换,得到符合要求的特征序列,构建新的距离公式对新的特征序列进行相似性度量。数值实验结果表明,与传统的动态时间弯曲度量方法相比,在分类效果和计算性能上具有一定的优势。通过实验发现,针对波动较为平缓的数据进行挖掘,将会取得更好的效果。特别地,在数据压缩量较大的低维空间下,新方法在保证分类质量的前提下,能够获得较好的时间效率。另外,参数选择是一个至关重要的步骤,不同参数对实验结果具有一定的影响,将来需要对参数选择做进一步探讨和研究。

参考文献:

[1]韩敏, 许美玲, 王新迎. 多元时间序列的子空间回声状态网络预测模型[J]. 计算机学报, 2014, 37(11): 2268-2275.
HAN Min, XU Meiling, WANG Xinying. A multivariate time series prediction model based on subspace echo state network[J]. Chinese journal of computers, 2014, 37(11): 2268-2275.

[2]MARSZA EK A, BURCZY SKI T. Modeling and forecasting financial time series with ordered fuzzy candlesticks[J]. Information sciences, 2014, 273: 144-155.

[3]ZAMORA M, LAMBERT A, MONTERO G. Effect of some meteorological phenomena on the wind potential of Baja California[J]. Energy procedia, 2014, 57: 1327-1336.

[4]GRAVIO G D, MANCINI M, PATRIARCA R, et al. Overall safety performance of air traffic management system: forecasting and monitoring[J]. Safety science, 2015, 72: 351-362.

[5]李海林, 郭韧 万校基. 基于特征矩阵的多元时间序列最小距离度量方法[J]. 智能系统学报, 2015, 10(3): 442-447.
LI Hailin, GUO Ren, WAN Xiaoji. A minimum distance measurement method for multivariate time series based on the feature matrix[J]. CAAI transactions on intelligent systems, 2015, 10(3): 442-447.

[6]SAKOE H, CHIBA S. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE transactions on acoustics, speech, and signal processing, 1978, 26(1): 43-49.

[7]IZAKIAN H, PEDRYCZ W, JAMAL I. Fuzzy clustering of time series data using dynamic time warping distance[J]. Engineering applications of artificial intelligence, 2015, 39: 235-244.

[8]ZHANG Zheng, TANG Ping, DUAN Rubing. Dynamic time warping under pointwise shape context[J]. Information sciences, 2015, 315: 88-101.

[9]李正欣, 张凤鸣, 李克武. 基于 DTW 的多元时间序列模式匹配方法[J]. 模式识别与人工智能, 2011, 24(3): 425-430.
LI Zhengxin, ZHANG Fengming, LI Kewu. DTW based pattern matching method for multivariate time series[J]. Pattern recognition and artificial intelligence, 2011, 24(3): 425-430.

[10]LI Hailin. Asynchronism-based principal component analysis for time series data mining[J]. Expert systems with applications, 2014, 41(6): 2842-2850.

[11]KEOGH E, PAZZANI M J. Derivative dynamic time warping[C]//Proceedings of the 1st SIAM International Conference on Data Mining. Chicago, IL, USA, 2001: 1-11.

[12]JEONG Y S, JAYARAMAN R. Support vector-based algorithms with weighted dynamic time warping kernel function for time series classification[J]. Knowledge-based systems, 2015, 75: 184-191.

[13]GÓRECKI T, ŁUCZAK M. Using derivatives in time series classification[J]. Data mining and knowledge discovery, 2013, 26(2): 310-331.

[14]李海林, 杨丽彬. 时间序列数据降维和特征表示方法

[J]. 控制与决策, 2013, 28(11): 1718-1722.
LI Hailin, YANG Libin. Method of dimensionality reduction and feature representation for time series[J]. Control and decision, 2013, 28(11): 1718-1722.

[15] BANKÓ Z, ABONYI J. Mixed dissimilarity measure for piecewise linear approximation based time series applications[J]. Expert systems with applications, 2015, 42(21): 7664-7675.

[16] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient similarity search in sequence databases[C]//Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms. Berlin Heidelberg: Springer, 1993, 730: 69-84.

[17] WANG Hongfa. Clustering of hydrological time series based on discrete wavelet transform[J]. Physics procedia, 2012, 25: 1966-1972.

[18] WENG Xiaoqing, SHEN Junyi. Classification of multivariate time series using two-dimensional singular value decomposition[J]. Knowledge-based systems, 2008, 21(7): 535-539.

[19] LIN J, KEOGH E, WEI Li, et al. Experiencing SAX: a novel symbolic representation of time series[J]. Data mining and knowledge discovery, 2008, 15(2): 107-144.

[20] KEOGH E, XI X, WEI L, et al. The UCR time series classification/clustering homepage [EB/OL]. 2007. http://www.cs.ucr.edu/~eamonn/time_series_data/.

作者简介：

李海林，男，1982 年生，副教授，博士，主要研究方向为数据挖掘与决策支持，主持国家自然科学基金、省部级基金多项，发表学术论文 30 余篇，其中被 SCI 检索 11 篇，EI 检索 10 余篇。

梁叶，女，1992 年生，硕士研究生，主要研究方向为数据挖掘与金融数据分析。

2016 年第 11 届 IEEE 国际应用计算智能和信息研讨会
2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI)

Authors are welcome to submit original and unpublished papers and attend the IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI 2016) to be held on May 12-14, 2016 in Timisoara, Romania.

TOPICS include but not limited to
Computational Intelligence
Intelligent Mechatronics
Systems Engineering
Intelligent Manufacturing Systems
Intelligent Control
Intelligent Robotics
Informatics
Website: http://conf.uni-obuda.hu/saci2016/