文章编号: 1001-0920(2017)03-0451-08 DOI: 10.13195/j.kzyjc.2016.0326

基于数值符号和形态特征的时间序列相似性度量方法

李海林†, 梁 叶

(1. 华侨大学 工商管理学院，福建 厦门 361021；2. 华侨大学 应用统计与大数据研究中心，福建 厦门 361021)

摘 要: 针对传统符号聚合近似方法在特征表示时容易忽略时间序列局部形态特征的局限性, 以及动态时间弯曲在度量上的优势, 提出一种基于数值符号和形态特征的时间序列相似性度量方法. 将时间序列进行符号和形态的特征表示后, 提出动态时间弯曲与符号距离结合的时间序列距离度量方法, 使所提方法能够较好地反映时间序列数据数值分布和形态特征. 实验结果表明, 所提出的方法在时间序列数据挖掘中能够得到较好的分类效果, 具有一定的优越性.

关键词: 符号聚合近似；动态时间弯曲；数值符号；形态特征；相似性度量

中图分类号: TP273 **文献标志码:** A

Similarity measure based on numerical symbolic and shape feature for time series

LI Hai-lin†, LIANG Ye

(1. School of Business Management，Huaqiao University，Xiamen 361021，China；2. Research Center for Applied Statistics and Big Data，Huaqiao University，Xiamen 361021，China)

Abstract: In view of the limitations of traditional symbolic aggregate approximation which easily neglect the local morphological character of time series when applied to feature representation, and the advantages of dynamic time warping in the similarity measurement, a similarity measure based on the numeric symbolic and shape feature is proposed. Through representing based on the numeric symbolic and shape feature, a distance measure combined with dynamic time warping and symbolic distance measure is proposed, which has a better reflection of data distribution and morphologic characteristic. Experimental results show that the proposed method has better effect of classification and certain superiority.

Keywords: symbolic aggregate approximation；dynamic time warping；numeric symbolic；shape feature；similarity measure

0 引 言

时间序列是一种生活中常见的数据形式, 在时间序列数据挖掘领域中, 特征表示和相似性度量是近十几年来的研究热点. 时间序列数据的复杂性易影响其挖掘的效率和质量, 为了提高其挖掘的技术性能, 研究者们通常先降低时间序列数据模型的复杂性, 利用数据降维或特征表示来清除冗余的数据信息, 以提高挖掘工作的效率和准确性[1]. 目前, 涌现出了很多特征表示方法, 如分段线性表示[2]、奇异值分解[3]、离散傅里叶变换[4] 和符号聚合近似[5] 等.

符号聚合近似 (SAX) 作为时间序列数据挖掘中一种常用的工具, 是一种高效的时间序列特征表示方法[6]. 它可以有效地处理时间序列的噪声, 压缩数据以提高数据的处理效率. SAX 的度量距离满足真实距离的下界要求, 保证了其距离值具有意义. 由于 SAX 简便、高效的特性, 在数据挖掘领域中得到了广泛的应用[7]. 然而, SAX 在运用过程中具有一定的局限性, 如分段数对挖掘结果的影响, 以及字符对刻画分段信息的完备性. 有学者利用分段均值和方差来描述分段特征[8], 但是当分段均值及方差都相等时, 该方法不能很好地描述分段的信息; 又有学者通过符号化拟合分段的斜率来描述分段具体形态特征[9],

收稿日期: 2016-03-20；**修回日期:** 2016-06-14.

基金项目: 国家自然科学基金项目 (61300139)；福建省社会科学规划项目 (FJ2016B076)；福建省自然科学基金项目 (2015J01581)；福建省高等学校新世纪优秀人才支持计划项目 (Z1625112).

作者简介: 李海林 (1982−), 男, 副教授, 博士, 从事数据挖掘与决策支持等研究；梁叶 (1992−), 女, 硕士生, 从事数据挖掘与金融数据分析的研究.

†通讯作者. E-mail: hailin@mail.dlut.edu.cn

但该方法是通过最小二乘法来拟合分段, 当压缩比较大且序列波动大时, 拟合的效果可能不佳; 有学者引入分段的起点和终点来构建新的特征序列[10], 但是当序列的分段起点和终点都与均值相等时, 则难以体现该分段的趋势. 另外, SAX 的符号数量对描述时间序列信息的完备性存在一定的影响.

动态时间弯曲 (DTW) 是最常用、灵活的度量方法, 最早应用于语音识别领域[11], 它可以在时域和频域捕获时间序列的多重特征[12], 也能够克服欧氏距离“点对点”的度量方式来达到度量不等长时间序列相似性的目的. 由于 DTW 允许时间轴弯曲, 通过弯曲时间轴来匹配数据点可以有效地根据时间序列的形态来进行度量. 尽管 DTW 可以通过弯曲时间和相位来得到高精确度的度量效果, 但它的高计算复杂度使其在大规模数据集的应用受到很大的限制, 使得需要基于特征表示的动态时间弯曲相似性度量方法[13-14].

鉴于以上 SAX 和 DTW 存在的问题, 为同时考虑时间序列数据具有数值分布差异和形态波动特征, 本文希望能够充分利用 SAX 在特征表示领域的优点和 DTW 对形态特征度量的优势, 提出基于数值符号和形态特征的相似性度量方法. 该方法利用 SAX 将时间序列进行分段且特征表示, 同时将分段以导数形式构建导数序列. 新方法结合 SAX 度量距离和 DTW 距离得到新方法下的相似性度量距离, 通过对两种特征序列的度量来达到对原始序列度量的目的. 分类实验结果表明, 与传统方法相比, 新方法能够较好地反映时间序列数据之间的数值差异和形态差异, 具有一定的优越性.

1 相关理论基础

1.1 符号聚合近似 (SAX)

SAX 共分为两个阶段. 首先, 将时间序列进行分段聚合, 将长度为 n 的序列 $S=(s_1,s_2,\cdots,s_n)$ 转化为另一条长度为 m 的时间序列 $Q=(q_1,q_2,\cdots,q_m)$, 实现时间序列的数据降维. 其中 $n>m$, 且令压缩比 $t=n/m$. 新序列 Q 中的任意元素为

$$q_i=\sum_{j=t\times(i-1)+1}^{t\times i} s_i,\ 1\leqslant i\leqslant m. \tag{1}$$

然后, 根据数据分布的概率来对分布空间进行划分, 每个分布空间用一种字符表示. 分段均值落在某个分布空间则用该分布空间所代表的字符表示, 最终实现时间序列的符号化特征表示. 不仅能够较好地体现时间序列的整体趋势, 有效地消除数据噪声, 而且还能够通过对数据压缩来提高运算效率.

当计算两条时间序列的符号化距离时, 服从高斯分布且有序的概率数值列表“分段点” $\beta_1,\beta_2,\cdots,\beta_{\alpha-1}$ 将时间序列分布空间划分为 $2,3,\cdots,\alpha$ 个等概率区域, 且 $\beta_{i+1}-\beta_i=\dfrac{1}{\alpha}$. 分段均值落在某个区域, 则由表示该区域的字符进行表征. 若原始时间序列 $S=(s_1,s_2,\cdots,s_n)$ 和 $Q=(q_1,q_2,\cdots,q_m)$ 转化为符号序列 $\widehat{S}=(\widehat{s}_1,\widehat{s}_2,\cdots,\widehat{s}_m)$ 和 $\widehat{Q}=(\widehat{q}_1,\widehat{q}_2,\cdots,\widehat{q}_m)$, 其中 $m<n$, 则其符号序列之间的距离为

$$\mathrm{DIST}(\widehat{S},\widehat{Q})=\sqrt{\frac{n}{k}}\cdot\sqrt{\sum_{i=1}^{k}\mathrm{dist}(\widehat{s}_i,\widehat{q}_i)^2}, \tag{2}$$

$\mathrm{dist}(\widehat{s}_i,\widehat{q}_i)^2$ 通过查找表 1 所给的数据之后利用下式计算得到:

$$\mathrm{dist}(\widehat{s}_i,\widehat{q}_i)=\begin{cases}0,\ |\widehat{s}_i-\widehat{q}_i|\leqslant 1;\\ \beta_{\max(\widehat{s}_i,\widehat{q}_i)-1}-\beta_{\min(\widehat{s}_i,\widehat{q}_i)},\ \text{otherwise.}\end{cases} \tag{3}$$

表 1 分段点及 3～10 个字符数查找表

β_i	3	4	5	6	7	8	9	10
β_1	−0.43	−0.67	−0.84	−0.97	−1.07	−1.15	−1.22	−1.28
β_2	0.43	0	−0.25	−0.43	−0.57	−0.67	−0.76	−0.84
β_3	–	0.67	0.25	0	−0.18	−0.32	−0.43	−0.52
β_4	–	–	0.84	0.43	0.18	0	−0.14	−0.25
β_5	–	–	–	0.97	0.57	0.32	0.14	0
β_6	–	–	–	–	1.07	0.67	0.43	0.25
β_7	–	–	–	–	–	1.15	0.76	0.52
β_8	–	–	–	–	–	–	1.22	0.84
β_9	–	–	–	–	–	–	–	1.28

1.2 动态时间弯曲 (DTW)

给定两条时间序列 $S=(s_1,s_2,\cdots,s_n)$ 和 $Q=(q_1,q_2,\cdots,q_m)$, 构建 $n\times m$ 的距离矩阵 $D_{n\times m}$, 其中

$$D(i,j)=\sqrt{(s_i-q_j)^2} \tag{4}$$

为点 s_i 与点 q_j 之间的欧氏距离. DTW 的目的在于找到一条最优弯曲路径 $P=(p_1,p_2,\cdots,p_K)$, 使得序列 S 和 Q 的全局代价 (累积距离)

$$\mathrm{DTW}(S,Q)=\min_P\Big(\frac{1}{K}\sum_{k=1}^{K}p_k\Big),\ k=1,2,\cdots,K \tag{5}$$

最小, 其中 p_k 表示该路径元素在距离矩阵中的位置, 且通常需要满足边界性、连续性和单调性[11], 有 $p_1=(1,1),p_K=(n,m)$. 有效的弯曲路径一般包含在累积矩阵中, 利用距离矩阵 $D_{n\times m}$, 通过动态规划方法来求解累积矩阵 $\gamma_{n\times m}$, 其中

$$\gamma(i,j)=D(i,j)+\min\begin{cases}\gamma(i-1,j-1),\\ \gamma(i-1,j),\\ \gamma(i,j-1).\end{cases} \tag{6}$$

可知, 两条时间序列的弯曲距离为 $\mathrm{DTW}(S,Q) = \gamma(n,m)$. 最后, 再反向以 p_K 为起点来寻找弯曲路径, 寻找公式如下:

$$p_{k-1} = \begin{cases} (1, j-1), & i = 1; \\ (i-1, 1), & j = 1; \\ \operatorname{argmin} \begin{cases} \gamma(i-1, j-1), \\ \gamma(i-1, j), \\ \gamma(i, j-1), \end{cases} & \text{otherwise.} \end{cases} \tag{7}$$

直到 $i = j = 1$ 及 $p_k = (1,1)$ 时, 弯曲路径寻找过程结束. 为了缓解过度弯曲, 在式 (7) 中, 当 $\gamma(i-1,j-1)$ 与 $\gamma(i-1,j)$ 或 $\gamma(i,j-1)$ 相等时, 取 $\gamma(i-1,j-1)$ 作为弯曲路径代价计算的贡献元素.

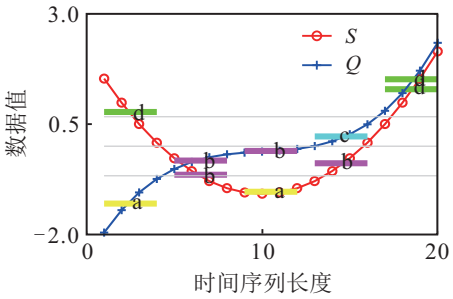
2　特征表示及相似性度量

传统 SAX 是一种常用的特征表示方法, 它通过将时间序列进行分段, 以字符化分段均值来表达时间序列整体特性, 且满足真实距离的下界要求. 然而, SAX 只是利用了分段的均值, 难免存在一定的局限性. 因此, 本节首先对 SAX 的局限性进行分析, 由此提出基于数值符号和形态特征的时间序列相似性度量新方法.

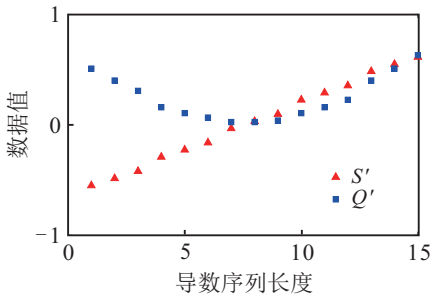
2.1　局限性分析

由于 SAX 对时间序列进行等分, 以符号化均值来表征分段信息, 不仅达到了数据降维的目的, 还近似地反映了时间序列的整体波动趋势.

如图 1(a) 所示, SAX 能够反映原始时间序列数据在分段后的整体信息, 并且通过字符串模式来匹配两条时间序列的具体关系. S 和 Q 分别通过 SAX 特征表示之后, 转换成字符串“d,b,a,b,d”和“a,b,b,c,d”, 通过式 (2) 计算字符串之间的相似性. 尽管 SAX 可以反映时间序列的整体特征, 但是仅仅利用符号化的均值来衡量分段的特征容易出现遗漏时间序列局部形态特征的问题. 例如, 出现分段均值一样但波动方向或者波动振幅不同等情况时, 传统方法不能如实地反映出来.



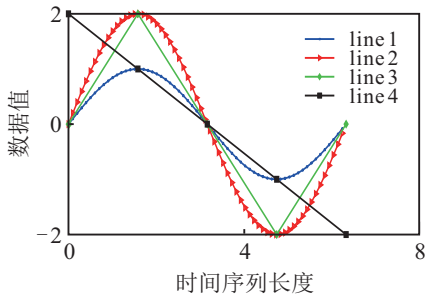
(a) 传统SAX表示方法



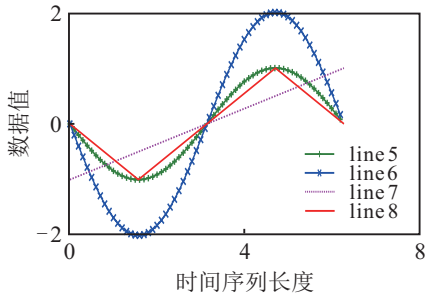
(b) 分段导数序列

图 1　传统 SAX 表示方法及分段导数序列

图 2 显示了两组共 8 条局部趋势明显不同但均值都相等的分段序列, 仅使用传统 SAX 来刻画分段信息, 将不能很好地衡量局部形态的差异性, 因此 SAX 具有一定的局限性. 文献 [10] 对 SAX 进行了改进, 提出 SAX-TD 方法. 以分段起点和终点分别与均值的差值来构建分段趋势, 并结合该分段映射的字符构建趋势符号序列. 该方法具有一定的优越性, 但如图 2 中出现的情况, 若分段起点和终点与均值一致时, 则该方法分段序列中表示趋势的数值为 0, 未能很好地体现局部的特征.



(a) 线条组1



(b) 线条组2

图 2　形态趋势不同而均值相同的两组线条

2.2　特征表示

本文从细化分段的局部特征, 不局限于仅依靠某些点来反映分段趋势的角度出发, 希望通过一种简单而有效的方式来刻画分段的上升、下降、凹、凸等形态特征. 而数值导数在刻画序列上升下降等趋势时既简单又高效, 若导数为正数, 则表示为上升; 若为负数, 则表示下降; 0 则表示无变化. 在计算时间序列导数时, 可以通过下式求解:

$$s'_i = s_i - s_{i-1}. \tag{8}$$

可以发现, 式 (8) 简单便捷, 由此得到导数序列的正负值能够充分考虑分段的局部趋势信息.

如图 1(b) 所示, S' 和 Q' 分别为图 1 中时间序列 S 和 Q 的导数序列. 时间序列 S 通过 SAX 被分为 5 个分段, 每个分段包含 4 个数据点, 并且每个分段通过式 (8) 可以得到 3 个导数值. 最终使得 S 的 20 个数据点转变为 S' 的 15 个导数数值, S' 则由 5 个分段、每个分段 3 个数据值构成. 容易发现, 导数序列中的具体数据能够很好地反映原始时间序列中具有局部近似数据值的形态差异性, 同时可以根据导数序列中具体数值的大小变化来反映时间序列趋势变化的快慢. 例如, 序列 S 和 Q 的第二分段均被字符“b”表征, 但是第二分段的导数序列 S' 的为负而 Q' 的为正, 所以该分段趋势也不同. 因此, 在研究符号化特征表示时, 不仅可以继续利用传统 SAX 来反映原始时间序列具体数值之间的差异和全局波动信息, 还需要利用分段导数来衡量局部的形态特征.

由以上分析可知, 利用数值分布差异性的优势以及导数具体刻画形态特征的优点, 在相似性度量之前先对原始时间序列进行两种形式的特征表示. 利用 SAX 将原始时间序列 $S=(s_1,s_2,\cdots,s_n)$ 转换成符号序列 $\widehat{S}=(\widehat{s}_1,\widehat{s}_2,\cdots,\widehat{s}_k)$. 同时, 利用式 (8) 将对应的每个分段转换成分段导数序列 $S'=\{S'_1, S'_2,\cdots,S'_k\}$, 其中 S'_i 为第 i 个分段序列的导数序列.

2.3 相似性度量

在符号聚合近似度量中引进分段导数, 可以克服仅依靠分段均值而忽略其他重要信息的缺陷, 但其应用效果还依赖于距离函数的度量质量. 为了使新方法具有更好的特征表示和相似性度量效果, 对导数序列的度量还需要利用精确的度量方法. 由于 DTW 可以弯曲时间轴来匹配点和点, 根据形态来精确地度量时间序列, 因此利用 DTW 来度量分段的导数序列, 找到趋势相似的分段导数序列, 以此来增强 SAX 度量效果.

鉴于 SAX 对整体信息反映的充分性, 分段局部形态特征的重要性及 DTW 的度量优越性, 提出一种基于数值符号和形态特征的相似性度量方法 (NSM). 该方法能够综合考虑 SAX 特征表示的优势以及 DTW 对形态匹配的精确性, 具有更完善时间序列的度量效果.

为了能够得到时间序列整体形态在低维空间的特征表示, 对于时间序列 $S=(s_1,s_2,\cdots,s_n)$ 和 $Q=(q_1,q_2,\cdots,q_n)$, 通过分段聚合近似 (PAA) 方法对其进行等长度分段, 得到 k 个分段之后计算它们的均值并将其符号化表示, 即 $\widehat{S}$ 和 $\widehat{Q}$, 达到数据降维和特征表示的目的. 同时, 由于分段导数可以充分体现时间序列的局部波动趋势特点, 对每个分段进行数值求导, 得到 k 个分段导数序列 $S'=\{S'_1,S'_2,\cdots,S'_k\}$ 和 $Q'=\{Q'_1,Q'_2,\cdots,Q'_k\}$. 另外, DTW 可以精确度量两条导数序列, 以很好地匹配两者的趋势形态. 因此, 综合考虑数据分布差异性和形态波动特征的相关性, 提出反映特征序列数据信息的距离度量方法, 即

$$\mathrm{Dist}(S,Q)=\sqrt{\frac{n}{k}}\cdot\sqrt{\sum_{i=1}^{k}\Big(\mathrm{dist}(s_i,q_i)^2+\frac{k}{n}\mathrm{DTW}(S'_i,Q'_i)\Big)}. \quad (9)$$

结合上述思想, 本文提出的基于数值符号和形态特征的相似性度量方法步骤如下.

输入: 时间序列 $S=(s_1,s_2,\cdots,s_n)$ 和 $Q=(q_1,q_2,\cdots,q_n)$, 分段数目 k 以及参数 α;

输出: 度量距离 $\mathrm{Dist}_{\mathrm{NSM}}(S,Q)$.

Step 1: 将时间序列 S 和 Q 平均划分为 k 段, 求出每个分段均值, 并将分段均值分别映射到由 $\alpha-1$ 个分布空间所表示的字符, 得到时间序列符号化序列, 即 $\widehat{S}=(\widehat{s}_1,\widehat{s}_2,\cdots,\widehat{s}_k)$ 和 $\widehat{Q}=(\widehat{q}_1,\widehat{q}_2,\cdots,\widehat{q}_k)$. 每个子序列长度分别为 $w=n/k$.

Step 2: 对分段的时间序列 S 和 Q, 利用式 (8) 分别求解各自每个分段的导数序列 S'_i 和 Q'_i, 最终得到 k 个分段导数序列 S' 和 Q'.

Step 3: 利用符号化序列 S 与 Q 和分段的导数序列 S' 与 Q', 根据式 (9) 求解时间序列 S 和 Q 的距离.

2.4 时间复杂度

根据式 (9) 进行时间复杂度分析, 同等长度为 n 的时间序列进行符号聚合近似, 其时间复杂度为 $O(n)$, 计算其 k 个分段一阶导数的时间复杂度为 $O(n)$, 利用 DTW 对 k 分段导数序列进行度量的时间复杂度为 $O\Big(k\Big(\frac{n}{k}-1\Big)^2\Big)$. 故新方法的整个特征表示和距离度量过程所需要的时间复杂度为 $O\Big(\frac{n^2}{k}+2n+k\Big)$. 然而, 传统动态时间弯曲的时间复杂度为 $O(n^2)$, 且又有 $n\gg 2$ 和 $k\ll n$, 通过 NSM 时间复杂度与 DTW 时间复杂度之间大小关系分析可知

$$\frac{n^2/k+2n+k}{n^2}=\frac{1}{k}+\frac{2}{n}+\frac{k}{n^2},$$

故 NSM 时间复杂度比动态时间弯曲的时间复杂度要小. 另外, 当 $k=1$ 时, 新方法的时间复杂度趋近于 DTW 所需要的时间 $O(n^2)$; 当 $k=n$ 时, 新方法的时间复杂度与传统 SAX 方法所需要时间近似, 即 $O(n)$.

3　数值实验

为了更好地分析 NSM 的度量效果, 采用传统的欧氏距离、动态时间弯曲、分段聚合近似、符号聚合近似、SAX-TD 以及 Górecki 和 Łuczak 提出的 DD_{DTW} 方法[15] 进行分类结果对比实验, 以验证本文方法的有效性及其时间效率.

3.1　数 据 集

采用 Keogh 教授提供的 38 个数据集[16] 进行实验, 每一个数据集包含训练集和测试集, 并且数据类别的范围为 2 到 50, 时间序列长度从 24 到 900 不等. 此外, 这些数据集类型包含了人工数据、现实数据以及图形数据, 由此可以通过不同类型的数据发现不同方法度量效果的差异. 由于每个数据集的时间序列取值范围不同, 存在的量纲会对时间序列的度量造成一定影响, 因此在度量之前需要对其进行归一化处理, 使其观测值服从高斯分布, 即均值为 0, 方差为 1. 表 2 给出了该 38 个数据集的基本信息, 包括数据集名称、类别个数 (C.N)、序列长度 (Len.)、训练集序列个数 (Tr.N) 和测试集序列个数 (Test.N). 数据集名 D.S.R., E.F.D., I.P.D., M.I., M.P.O.A.G., M.P.O.C., M.P.TW, S.A.R.S., S.A.R.S.II, S.C, T.S.1, T.S.2, T.L.E., W.S. 分别表示 Diatom Size Reduction, ECG Five Days, Italy Power Demand, Medical Images, Middle Phalanx Outline Age Group, Middle Phalanx Outline Correct, Middle Phalanx TW, Sony AIBO Robot Surface, Sony AIBO Robot Surface II, Synthetic Control, Toe Segmentation1, Toe Segmentation2, Two Lead ECG, Words Synonyms.

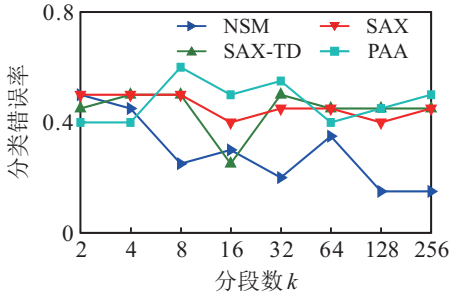
表 2　数据集基本信息

名称	C.N	Len.	Tr.N	Test.N
50words	50	270	450	455
Adiac	37	176	390	391
Arrow Head	3	251	36	175
Beef	5	470	30	30
Beetle Fly	2	512	20	20
Bird Chicken	2	512	20	20
Car	4	577	60	60
CBF	3	128	30	900
Coffee	2	286	28	28
Computer	2	720	250	250
D. S. R.	4	345	16	306
Earthquake	2	512	139	322
ECG200	2	96	100	100
E. F. D.	2	136	23	861
Face Four	4	350	24	88
Gun Point	2	150	50	150
Herring	2	512	64	64
I. P. D.	2	64	67	1 029
Lighting2	2	637	60	61
Lighting7	7	319	70	73
Meat	3	448	60	60
M. I.	10	99	381	760
M. P. O. A. G.	3	80	154	400
M. P. O. C.	2	80	291	600
M. P. TW	6	80	154	399
Mote Strain	2	84	20	1 252
Olive Oil	3	570	30	30
plane	7	144	105	105
ShapeletSim	2	500	20	180
S.A.R.S.	2	70	27	953
S.A.R.S.II	2	65	20	601
S.C.	6	60	300	300
T.S.1	2	277	40	228
T.S.2	2	343	36	130
T.L.E.	2	82	23	1 139
Wine	2	234	57	54
W.S.	25	270	267	638
worms	5	900	77	181

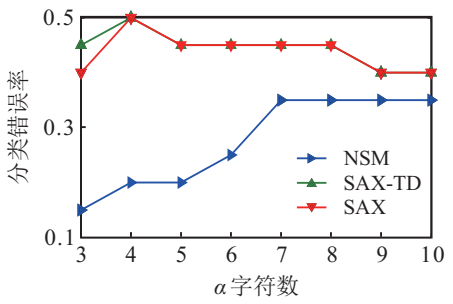
3.2　分类实验

由于分类效果的好坏取决于度量距离效果的好坏, 实验采用最近邻 (1NN) 分类方法直接对测试集进行分类, 以此观察本文方法与其他方法度量效果的差别. 给定分段数目范围为 $[2, n/2]$, 且步长为 1 到 k, n 为时间序列长度. 给定字符数 $\alpha = 3, 4, \cdots, 10$, 观察每种分段数在 α 取值不同情况下的分类情况. 实验取使分类错误率最小的 k 以及 α 作为度量距离的最佳分段长度及字符数, 当出现多个相同分类错误率的情况时, 取最小的 k 及 α 值.

图 3 显示了数据集 Birdchiken 的分段数 k 和字符数 α 不同取值时, 分类错误的比较情况. 图 3(a) 为当字符数固定为 4, 分段数 k 取值为 2 到 256 的错误率情况. 容易发现, NSM 方法出现随着分段数取值的不断增长, 错误率取值呈现下降的趋势. 这表



(a) 分段数对错误率的影响



(b) 字符数对错误率的影响

图 3　不同 k、α 取值对分类错误率的影响

明分段数越多, 所表征时间序列的信息越详细. 另外, SAX-TD 方法与 SAX 方法结果取值波动情况类似, 可能针对某些数据集不一定能够得到改进的分类效果. 图 3(b) 为固定分段数 k 为 128 时, α 取值从 3 到 10 时错误率的分布情况, 易知新方法 NSM 能够获得较好的分类效果. 由于 PAA 无需 α 参数, 因此不参与比较. 有学者[17] 指出, 参数 α 取 3 或 4 对几乎任何数据集的挖掘的效果都是较为合适的. 从图中也可以发现, 尽管 α 的不同取值对分类错误率有一定的影响, 但是影响是较为平缓的.

表 3 给出了不同数据集的分类实验结果, 以及 NSM 在取得最小分类错误率的最小分段数 k 及 α 值. 底部对平均分类错误率、标准差及分类胜出率进行统计比较, 其中胜出率为该方法下得到分类错误率最小的数据集个数与实验数据集总个数的比值. 从表 3 的结果比较易知, NSM 得到了更小的平均分类错误率, 且实验结果的标准差也是最低的, 说明 NSM 在分类精度上具有较为稳定的优势. $\mathrm{DD_{DTW}}$ 方法结合了原始时间序列 DTW 距离与一阶导数序列 DTW 距离, 在一定程度上提高了分类精度. 尽管 NSM 的分类胜出率低于 $\mathrm{DD_{DTW}}$, 但根据 $\mathrm{DD_{DTW}}$ 的分类结果也可以从另一方面说明, 利用时间序列的一阶导数有助于 NSM 刻画时间序列的形态特征, 甚至有些数据集的错误率呈现明显的降低, 如 Herring 数据集, ShapeletSim 数据集等. 通过将表 3 转化为图 4, 可以更为直观地呈现各种传统方法与新方法的分类结果差异性比较, 如图 4 所示.

表 3 分类错误率 %

名称	SAXTD	$\mathrm{DD_{DTW}}$	SAX	Euc	DTW	PAA	NSM(k,α)
50words	0.305 5	0.239 6	0.312 1	0.369 0	0.310 0	0.876 9	0.222 8(54,9)
Adiac	0.340 2	0.294 1	0.381 1	0.389 0	0.396 4	0.879 8	0.364 1(44,3)
Arrow Head	0.211 4	0.205 7	0.222 9	0.200 0	0.297 0	0.537 1	0.222 8(25,7)
Beef	0.233 3	0.433 3	0.300 0	0.333 0	0.367 0	0.400 0	0.300 0(10,6)
Beetle Fly	0.150 0	0.250 0	0.150 0	0.250 0	0.300 0	0.400 0	0.050 0(126,3)
Bird Chicken	0.250 0	0.250 0	0.300 0	0.450 0	0.250 0	0.400 0	0.150 0(128,3)
Car	0.200 0	0.200 0	0.216 7	0.267 0	0.267 0	0.600 0	0.216 6(24,9)
CBF	0.071 1	0.003 3	0.054 4	0.148 0	0.003 3	0.400 0	0.060 0(8,6)
Coffee	0.000 0	0.107 1	0.035 7	0.000 0	0.000 0	0.392 9	0.071 4(13,4)
Computer	0.424 0	0.348 0	0.452 0	0.424 0	0.300 0	0.388 0	0.330 0(120,10)
D.S.R.	0.019 6	0.032 7	0.026 1	0.065 0	0.032 7	0.274 5	0.022 8(23,6)
Earthquake	0.189 4	0.257 8	0.177 0	0.326 0	0.258 0	0.273 3	0.225 3(16,4)
ECG200	0.080 0	0.130 0	0.080 0	0.120 0	0.230 0	0.350 0	0.000 0(8,3)
E.F.D.	0.131 2	0.227 6	0.123 1	0.203 3	0.232 3	0.468 1	0.130 2(34,8)
Face Four	0.113 6	0.170 5	0.102 3	0.216 0	0.170 5	0.590 9	0.136 3(50,9)
Gun Point	0.106 7	0.013 3	0.120 0	0.087 0	0.093 3	0.366 7	0.066 7(75,4)
Herring	0.312 5	0.453 1	0.328 1	0.484 0	0.469 0	0.406 3	0.000 0(16,3)
I.P.D.	0.041 8	0.047 6	0.047 6	0.045 0	0.050 0	0.338 2	0.061 2(8,7)
Lighting2	0.229 5	0.098 4	0.196 7	0.246 0	0.131 1	0.442 6	0.147 5(49,6)
Lighting7	0.287 7	0.274 0	0.301 4	0.425 0	0.274 0	0.726 0	0.287 6(29,5)
Meat	0.000 0	0.066 7	0.000 0	0.067 0	0.067 0	0.200 0	0.000 0(14,5)
M.I.	0.325 0	0.256 6	0.340 8	0.316 0	0.263 2	0.615 8	0.331 6(33,9)
M.P.O.A.G	0.252 5	0.235 0	0.257 5	0.260 0	0.250 0	0.612 5	0.252 5(20,5)
M.P.O.C.	0.270 0	0.231 7	0.388 3	0.247 0	0.352 0	0.443 3	0.273 3(20,6)
M.P.TW	0.403 5	0.416 0	0.406 0	0.439 0	0.416 0	0.498 7	0.313 5(40,5)
Mote Strain	0.126 2	0.165 3	0.180 5	0.121 0	0.165 3	0.278 0	0.164 5(12,10)
Olive Oil	0.033 3	0.100 0	0.033 3	0.1330	0.167 0	0.366 7	0.066 7(10,6)
plane	0.019 0	0.000 0	0.028 6	0.038 0	0.000 0	0.514 3	0.000 0(48,3)
ShapeletSim	0.400 0	0.355 6	0.377 8	0.461 0	0.350 0	0.455 6	0.238 9(50,5)
S.A.R.S.	0.153 1	0.242 9	0.163 1	0.305 0	0.275 0	0.271 2	0.232 9(14,10)
S.A.R.S.II	0.124 9	0.104 9	0.192 0	0.141 0	0.168 9	0.237 1	0.238 2(5,9)
S.C.	0.060 0	0.006 7	0.006 7	0.120 0	0.007 0	0.333 3	0.080 0(10,8)
T.S.1	0.302 6	0.241 3	0.298 2	0.320 0	0.228 0	0.473 7	0.201 8(69,7)
T.S.2	0.100 0	0.169 2	0.100 0	0.192 0	0.162 0	0.376 9	0.107 6(9,6)
T.L.E.	0.189 6	0.007 9	0.195 8	0.253 0	0.096 0	0.411 8	0.117 6(41,10)
wine	0.203 7	0.425 9	0.259 3	0.389 0	0.426 0	0.222 2	0.166 7(3,9)
W.S.	0.362 1	0.264 9	0.360 5	0.382 0	0.351 1	0.841 7	0.260 2(135,10)
worms	0.541 4	0.519 3	0.552 5	0.635 0	0.536 0	0.629 8	0.536 0(100,4)
MEAN	0.189 3	0.206 9	0.203 1	0.249 5	0.221 0	0.438 5	**0.166 4**
STD	0.133 4	0.138 0	0.140 8	0.148 5	0.138 9	0172 3	**0.122 5**
胜出率	0.263 1	0.368 4	0.184 2	0.078 9	0.131 5	0	0.289 5

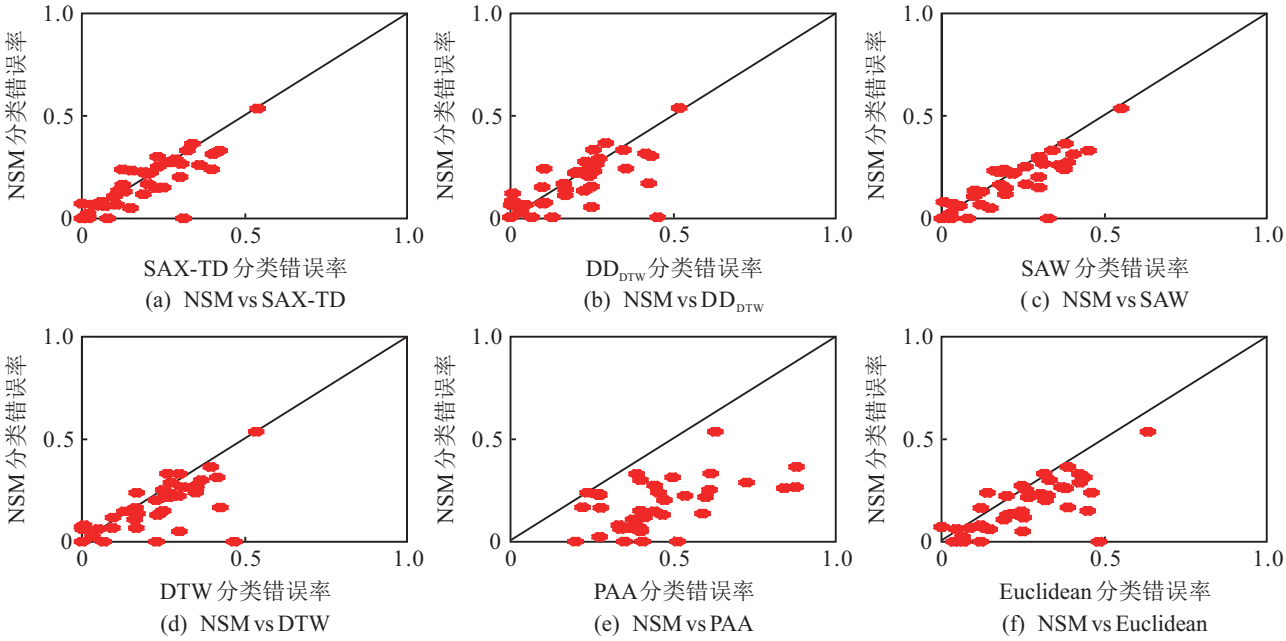


图 4　分类错误率比较

图 4 分别给出 NSM 与 SAX-TD、DD_{DTW}、DTW、SAX、PAA 及 uclidean 在分类中的错误率比较. 为了便于直观地比较, 将各个方法的分类错误率比较以散点图的形式展现出来. 错误率分布范围在 [0,1] 之间, 数据偏向的一方, 表示该一方的方法取得较差的分类效果. 可以发现, 在各类方法与 NSM 进行比较时, 大部分的错误率散点分布都偏离 NSM 方法, 甚至图 4(e) 中的散点几乎全部偏向右下方, 这说明了 NSM 方法在分类上能取得更好的效果.

3.3　复杂度分析

动态时间弯曲的局限在于它的高时间消耗性, 而 NSM 在分段中结合了 DTW 高度量精确性, 但难免也会受到其高时间消耗的影响. 因此, 选择合适的参数 k, 以使得 NSM 总体时间消耗小于直接使用 DTW 进行度量的时间消耗就显得尤为重要.

如图 5 所示, 根据 Birdchicken 数据集的分段情况将 NSM、SAX-TD、DD_{DTW}、DTW、SAX、PAA 以及欧氏距离的时间消耗展示出来. 从图 5(a) 可以发现, NSM 随着分段数的增加, 其时间消耗开始不断减少. 同时, 当分段数增加到一定程度时, NSM 时间消耗与传统 SAX、PAA 和欧氏距离方法很接近. 这是由于分段数的增大, 使每个分段所包含的数据点减少, 在利用 DTW 度量分段导数时就大大降低了时间消耗, 进而验证了上述针对 NSM 的时间复杂度分析. 从图 5(b) 的结果可以看出, 由 DD_{DTW} 在每次的度量中都需要计算两次 DTW 值, 其时间消耗约为传统 DTW 的两倍. 尽管 NSM 在分段数取值较小的情况

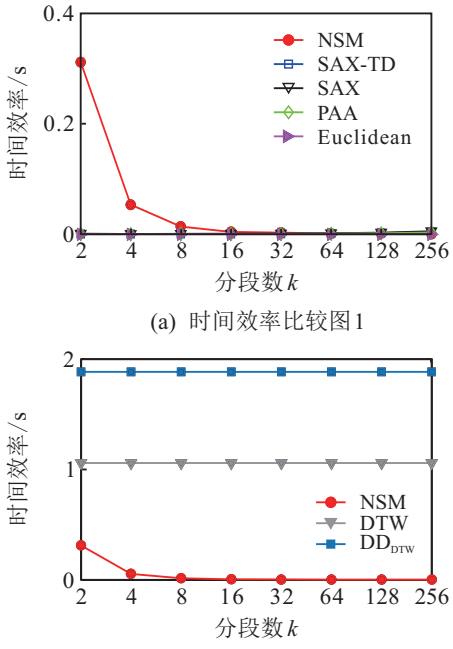


图 5　时间效率比较

下消耗的时间高于 SAX-TD、SAX 等方法, 但 NSM 的总体时间消耗仍然远低于 DD_{DTW} 和 DTW, 在一定程度上说明 NSM 在低维空间具有一定的高效性.

4　结　论

针对基于数据分布的符号化特征表示及其相似性度量存在容易忽略时间序列局部重要特征的情况, 本文同时从数据符号和形态特征的角度研究时间序列的相似性度量方法 (NSM), 以便更好地提高后续时间序列数据挖掘相关算法的性能和质量. 该方法在对时间序列进行符号聚合特征表示的同时, 对分

段数据求解一阶导数, 用分段的导数序列来表征分段的局部形态特征, 再用新度量公式对符号序列和导数序列进行计算, 实现基于新特征表示的时间序列相似性度量方法. 与传统方法相比, 新方法的优势体现在: 1) 通过 SAX 得到时间序列的符号特征序列来表示原始时间序列的总体信息, 不仅得到时间序列的整体波动情况, 还反映了时间序列的数据分布信息; 2) 时间序列经过分段之后, 每个分段利用一阶导数来描述时间序列局部的形态变化特征, 使得新特征表示方法能够较好地反映局部数据形态信息; 3) 动态时间弯曲可以实现“一对多”数据匹配, 因此新方法使得异步形态特征可以相互匹配; 4) 新方法利用 SAX 对数值分布的表述完整性、数值导数对分段形态特征描述的具体性以及动态时间弯曲的度量精确性, 综合考虑了数值分布差异性和形态波动特征, 提高了时间序列分类质量. 然而, 由于 NSM 结合了 DTW 方法, 在时间消耗上存在一定局限性, 这也是下一步要进行的研究.

参考文献 (References)

[1] 李海林. 基于变量相关性的多元时间序列特征表示 [J]. 控制与决策, 2015, 30(3): 441-447.
(Li H L. Feature representation of multivariate time series based on correlation among variables[J]. Control and Decision, 2015, 30(3): 441-447.)

[2] 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.
(Yan Q Y, Xia S X. An piecewise linear fitting algorithm for infinite time series[J]. Acta Electronica Sinica, 2010, 38(2): 443-448.)

[3] 吴虎胜, 张凤鸣, 钟斌. 基于二维奇异值分解的多元时间序列相似匹配方法 [J]. 电子与信息学报, 2014, 36(4): 847-854.
(Wu H S, Zhang F M, Zhong B. Similar pattern matching method for multivariate time series based on two-dimensional singular value decomposition[J]. J of Electronics and Information Technology, 2014, 36(4): 847-854.)

[4] 王伟, 刘国华, 徐斌. 不确定时间序列的降维及相似性匹配 [J]. 计算机科学与探索, 2015, 9(4): 418-428.
(Wang W, Liu G H, Xu B. Dimensionality reduction and similarity match of uncertain time series[J]. J of Frontiers of Computer Science and Technology, 2015, 9(4): 418-428.)

[5] Wan Y, Gong X, Si Y W. Effect of segmentation on financial time series pattern matching[J]. Applied Soft Computing, 2015, 38: 346-359.

[6] Rakthanmanon Q Z G B T, Keogh E. A novel approximation to dynamic time warping allows anytime clustering of massive time series datasets[C]. Proc of the 2012 SIAM Int Conf on Data Mining. California: SIAM, 2012: 999-1010.

[7] Georgoulas G, Karvelis P, Loutas T, et al. Rolling element bearings diagnostics using the symbolic aggregate approXimation[J]. Mechanical Systems and Signal Processing, 2015, 60: 229-242.

[8] 钟清流, 蔡自兴. 基于统计特征的时序数据符号化算法 [J]. 计算机学报, 2008, 31(10): 1857-1864.
(Zhong Q L, Cai Z X. The symbolic algorithm for time series data based on statistic feature[J]. Chinese J of Computers, 2008, 31(10): 1857-1864.)

[9] 李海林, 郭崇慧. 基于形态特征的时间序列符号聚合近似方法 [J]. 模式识别与人工智能, 2011, 24(5): 665-672.
(Li H L, Guo C H. Symbolic aggregate approximation based on shape features[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(5): 665-672.)

[10] Sun Y, Li J, Liu J, et al. An improvement of symbolic aggregate approximation distance measure for time series[J]. Neurocomputing, 2014, 138: 189-198.

[11] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE Trans on Acoustics, Speech and Signal Processing, 1978, 26(1): 43-49.

[12] Cai Q, Chen L, Sun J. Piecewise statistic approximation based similarity measure for time series[J]. Knowledge-Based Systems, 2015, 85: 181-195.

[13] Lin J, Keogh E, Wei L, et al. Experiencing SAX: A novel symbolic representation of time series[J]. Data Mining and Knowledge Discovery, 2007, 15(2): 107-144.

[14] 李正欣, 张凤鸣, 李克武. 多元时间序列模式匹配方法研究 [J]. 控制与决策, 2011, 26(4): 565-570.
(Li Z X, Zhang F M, Li K W. Research on pattern matching method for multivariate time series[J]. Control and Decision, 2011, 26(4): 565-570.)

[15] Górecki T, Łuczak M. Using derivatives in time series classification[J]. Data Mining and Knowledge Discovery, 2013, 26(2): 310-331.

[16] Chen Y P, Keogh E, Hu B, et al. The UCR time series classification archive[EB/OL]. (2015-07-01) [2015-12-01]. http://www.cs.ucr.edu/∼eamonn_ series _ data/.

[17] Tanaka Y, Uehara K. Motif discovery algorithm from motion data[C]. Proc of the 18th Annual Conf of the Japanese Society for Artificial Intelligence. Kanazawa: JSAI, 2004: 2-4.

（责任编辑：齐 霁）