一种基于始末距离的时间序列符号聚合近似表示方法

季海娟 周从华 刘志锋
(江苏大学计算机科学与通信工程学院 江苏 镇江 212013)

摘 要 时间序列数据的特征表示方法是时间序列数据挖掘任务的关键技术,符号聚合近似表示(SAX)是特征表示方法中比较常用的一种。针对SAX算法在各序列段表示符号一致时无法区分时间序列间的相似性这一缺陷,提出了一种基于始末距离的时间序列符号聚合近似表示方法(SAX_SM)。由于时间序列有很强的形态趋势,因此文中提出的方法选用起点和终点来表示各个序列段的形态特征,并使用各序列段的形态特征和表示符号来近似表示时间序列数据,以将其从高维空间映射到低维空间;然后,针对起点和终点构建始末距离来计算两序列段间的形态距离;最后,结合始末距离和符号距离定义一种新的距离度量方式,以更客观地度量时间序列间的相似性。理论分析表明,该距离度量满足下界定理。在20组UCR时间序列数据集上的实验表明,所提SAX_SM方法在13个数据集中获得了最高的分类准确率(包含并列最大的),而SAX只在6个数据集中获得了最高的分类准确率(包含并列最大的),因此SAX_SM具有比SAX更优的分类效果。

关键词 时间序列数据,序列段,始末距离,符号距离
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.039

Symbolic Aggregate Approximation Method of Time Series Based on Beginning and End Distance

JI Hai-juan ZHOU Cong-hua LIU Zhi-feng
(School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China)

Abstract The feature representation method of time series data is the key technology of time series data mining task, and the symbolic aggregate approximation (SAX) method is most commonly used in feature representation methods. A symbolic aggregate approximation method based on beginning and end distance (SAX_SM) was proposed because SAX algorithm can not distinguish the similarity between time series when the symbol is consistent in each sequence segment of time series. Time series data have a strong morphological trend, so the proposed method uses the beginning point and the end point to represent the morphological feature of each sequence segment, and then uses the morphological feature and representation symbol of each sequence segment to approximate the time series data, in order to map it from high-dimensional space to low-dimensional space. Next, in order to calculate the morphological distance between the two sequences, this paper constructed beginning and end distance based on the beginning point and the end point. Finally, to measure the similarity between time series more objectively, a new distance metric approach was defined by combining the beginning and end distance and the symbol distance. The theoretical analysis shows that the new distance measure satisfies the lower bound theorem. Experiments on 20 sets of UCR time series data sets show that the proposed SAX_SM method achieves the highest classification accuracy (including the largest side by side) in 13 data sets, while SAX only gets the largest classification accuracy in 6 data sets (including the largest side by side). Therefore, SAX_SM has better classification result than SAX.

Keywords Time series data, Sequence segment, Beginning and end distance, Symbol distance

1 引言

当今社会正处于一个复杂多样的数据时代,时间序列数据是其中一种重要的数据类型。一条时间序列由一组序列数据组成,这些数据一般是通过以相同的时间间隔对某个潜在过程进行采样选值而得到的。现实生活中,在一系列时间点上选取数据进行观测是一种普遍现象[1]。如在股市上,我们会观测股票每日的闭盘价、周利率;在气象上,我们会观测每日的最高气温和最低气温;在农业上,我们会观测不同农作物的年产量和年出/入口量等。随着社会经济和科学技术的发

到稿日期:2017-04-16 返修日期:2017-07-11 本文受江苏省重点研发计划(社会发展)项目(BE2016630),江苏省六大人才高峰项目(2014-WLW-012),江苏省重点研发计划(社会发展)项目(BE2015617),无锡市卫计委重点项目(Z201603)资助。
季海娟(1993-),女,硕士生,主要研究方向为大数据技术,E-mail:18260622771@163.com;周从华(1978-),男,博士,教授,主要研究方向为大数据技术、人工智能,E-mail:zchwyl2003@163.com(通信作者);刘志锋(1981-),男,博士,副教授,主要研究方向为大数据技术。

展,时间序列数据量迅猛增长。相应地,使用数据挖掘技术[2]在时间序列数据中“挖掘”出有意义的信息和模式已成为一个非常重要的问题,并吸引着越来越多研究者的关注[3]。在时间序列数据挖掘领域[4],时间序列通常由一系列的时间点构成,由于时间序列数据的时间跨度和时间间隔决定着其特征维度,使得时间序列数据的特征维度过多,直接对原始时间序列进行挖掘时效率较低,运行时间过长。此外,直接对具有过多特征维度的原始时间序列数据集进行相似性查询,会大大增加计算的复杂度。因此,亟需找到一种简单的时间序列表示方法和精确的距离度量方法,以提高时间序列数据挖掘的性能和效率。在大多数时间序列数据挖掘任务中,通常通过对时间序列数据集进行降维表示[5-7]这一预处理步骤,来简化时间序列的表示。

IBM 公司的 Pazzani 和 Agrawal 研究小组较早地开始了相关研究。1993 年,Agrawal 等人[8]首次提出了一种使用傅里叶变换(DFT)将时间序列的特征域变换到频域的时间序列表示方法,并将其运用于相似度查询中。该方法有效解决了时间序列挖掘中“特征抽取完备性”和“维度灾难”这两个问题,这一成果促使国内外研究者纷纷参与到时间序列的特征表示和相似度度量的研究工作中。美国加州大学的 Keogh 教授领导的研究小组在时间序列特征表示和相似度度量的研究中取得了丰硕的成果。Keogh 等人[13]提出了分段聚合近似表示方法(PAA),并通过在索引速度和灵活性等方面与传统的奇异值分解(SVD)和 DWT 进行比较,表明了其在时间序列相似性度量和索引上具有优势。国内在时间序列特征表示和相似度度量方面的研究起步较晚,研究者主要来自重点院校和各大研究所。

目前,学者们已提出了不少有效的时间序列表示方法。根据时间序列数据的不同转化方式,这些方法大体可分为 3 类:非数据适应性、数据适应性和基于模型的表示方法[9]。1)在非数据适应性表示方法中,每条时间序列的转换参数是一致的。例如,上文提到的 Agrawal 等人[8]提出的 DFT 表示方法就属于这一类方法,它有效地解决了“维数灾难”和“特征抽取完备”问题。自 DFT 表示方法被提出以来,欧氏距离也被广泛用于解决时间序列的相似性问题。2)数据适应性表示方法中,每条时间序列的转换参数是不一致的,该方法既依赖于单条时间序列,又受整体时间序列数据集的影响。例如,奇异值分解方法对时间序列数据集中的任一数据对象进行删除或者修改时,都会改变时间序列的表示结果。Lin 等人[10]提出了一种符号聚合近似表示方法(SAX),该方法可以将原时间序列从高维数据映射到符号表示的低维空间,但该方法在进行符号表示时只把握了时间序列的总体趋势而忽略了序列段之间的局部信息,尤其是当各个序列段的表示符号一致时,无法进行时间序列之间的相似性比较。3)基于模型的表示方法利用模型来定义时间序列变量间的关系,即先假设时间序列由某种模型产生,然后建立模型,最后利用该模型的参数来表示时间序列。例如,Azzouzil 等人提出的 HMM 模型[11]和 Kalpakis 等人提出的 ARIMA 模型[12]等都属于这一类方法。

SAX 方法[10]在各序列段表示符号一致时,无法进行时间序列之间的相似性比较。为了解决这一问题,本文提出了一种基于始末距离的时间序列符号聚合近似表示方法(SAX_SM)。由于时间序列有很强的形态趋势,因此本文提出的方法选用起点和终点来表示各个序列段的形态特征,并使用各序列段的形态特征和表示符号来近似表示时间序列数据,从而将其从高维空间映射到低维空间;然后,针对起点和终点构建始末距离,以计算两序列段间的形态距离;最后,结合始末距离和符号距离定义一种新的距离度量,以更客观地度量时间序列之间的相似性。理论分析证明,该距离度量满足下界定理。为了验证所提方法的有效性,在 20 组 UCR 时间序列数据集上进行了实验。实验结果表明,所提 SAX_SM 方法在 13 个数据集中获得了最高的分类准确率(包含并列最大的),而 SAX 只在 6 个数据集中获得了最高的分类准确率(包含并列最大的),这说明 SAX_SM 具有比 SAX 更优的分类性能。

本文第 2 节简要介绍时间序列数据表示方法的已有工作;第 3 节深入阐述基于始末距离的时间序列符号聚合近似表示方法,并提出一种新的距离度量公式;第 4 节给出理论分析和实验结果分析;最后总结全文并探讨下一步可能的研究方向。

2　相关工作

2.1　时间序列表示方法的相关研究

时间序列表示方法即是对时间序列数据集进行数据降维,对原时间序列数据集进行一定的转化,使其成为其他域中的近似表示序列,并使该近似表示序列能尽可能地反映原时间序列的信息。Lin 等人[10]提出的符号聚合近似(SAX)表示方法是以 Keogh 等人[13]提出的分段聚合近似(PAA)表示方法为基础的一种符号表示方法。该方法满足下界定理[14],但 SAX 是将分段的平均值进行符号化,会造成数据中其他信息的缺失,且只能反映原时间序列的总体变化趋势,而不能描述各段的局部信息,尤其当两个时间序列各段的均值一致而各段趋势不同时,SAX 的局限性更加明显。针对 SAX 存在的缺陷,越来越多的改进方法被提出。Lkhagva 等人[15]提出了一种扩展性的时间序列符号聚合近似表示方法(ESAX),该方法在使用序列段均值的基础上,还使用了最大值和最小值,但是其将最大值、最小值和均值放在同等地位进行比较的合理性有待商榷,而且距离的下界性也没有得到证明。钟清流等人[16]提出了一种基于统计特征方差的符号聚合近似表示方法,该方法使用均值的分割点对方差进行符号化,这样的计算方式缺乏合理性。Esmael[17]提出了一种融合斜率的符号聚合近似表示方法,该方法虽然考虑了序列段的内部特征(斜率),但是没有给出距离的度量公式。Sun[18]提出了一种基于趋势距离的符号聚合近似表示方法(SAX_TD),首先针对序列段的起点和终点构建趋势距离,然后集合趋势距离和 SAX 距离给出了新的距离度量。该方法在构建趋势距离时,增加了存储维度和运行时间。郑旭[19]提出了一种基于小波熵的时间序列聚合近似表示方法,该方法将小波熵运用于 PAA 算法的改进中,按各个区间内的小波熵值的比重分配各区间内的段数,但其在度量时没有提出新的距离度量公式,仍旧沿用原表示方法使用的欧氏距离进行度量。Wang[20]提出了一种新的距离度量公式,在表示符号的基础上引入了角度的概

念，但是该方法没有对时间序列数据进行降维，且由于在运行过程中调整了权重大小，时间复杂度较高。

2.2 传统的 SAX 算法

符号聚合近似表示方法的主要思想是：先使用分段聚合近似表示方法将时间序列进行分段，并获取各个序列段的均值，然后根据各个序列段的均值进行符号化赋值，由每个序列段所得符号集合对整条时间序列进行表示。不同的分段数目 ω 决定了近似表示后的维度，不同的字母表大小 α 决定了近似表示时字符的跨度。

使用 SAX 方法对时间序列 X 进行转换的步骤如下：

1)将时间序列 X 进行归一化和标准化，即将其标准化为正态分布均值为 0、标准差为 1 的标准序列 X'；

2)使用 PAA 算法进行降维，得到 X 的分段聚合近似表示序列$\bar{X}$；

3)对表示序列 $\bar{X}$ 进行符号化赋值。为了将表示序列 $\bar{X}$ 转化为 SAX 符号，定义“分割点”，即在高斯曲线下将分布空间划分成 α(字母表大小)个等概率的区域。在分割点的定义中，归一化的时间序列一般是符合高斯分布的[21]，即使因为有少数不符合高斯分布的时间序列数据使得效率略有恶化，但是算法的正确性不会受到影响。根据所选的字母表大小 α，计算每个序列段该被赋予的符号。由于$\bar{X}$近似服从高斯分布，因此可以将其近似划分成 α 个等概率的区间(划分区间的分割点 β_i 可以根据表 1 得到)，用相同的符号表示位于分割点 β_i 和 β_{i+1} 之间的各个序列段的均值$\overline{X_i}$。

表 1 字母表大小从 3 到 10 分别对应的分割点

Table 1 Corresponding split points about alphabet size from 3 to 10

β_i	α							
	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	0.67	0.14	0
β_6	—	—	—	—	1.07	0.32	0.43	0.25
β_7	—	—	—	—	—	1.15	0.76	0.52
β_8	—	—	—	—	—	—	1.22	0.84
β_9	—	—	—	—	—	—	—	1.28

3 基于始末距离的时间序列符号聚合近似表示方法

原符号聚合近似表示方法 SAX[10] 中只使用各序列段的表示符号来近似表示时间序列数据。每条时间序列数据都有其自己的形态趋势(上升、下降或者水平)，这便使得经过划分之后的时间序列的各个序列段也都有各自的形态趋势(上升、下降或者水平)。对形态趋势的度量有多种方法，本文提出的基于始末距离的时间序列符号聚合近似表示方法(SAX_SM)使用各个序列段的起点和终点作为该序列段的形态特征，以度量各个序列段的形态趋势。起点和终点在时间序列形态分析中有着重要的作用，不同的形态趋势有不同的起点和终点，上升形态的起点值肯定低于终点值，下降形态的起点值肯定高于终点值，水平形态的起点值肯定与终点值相差不大，这便使得在时间序列各个序列段中使用起点和终点来表示各序列段的形态特征具有普遍适用性。SAX_SM 方法结合各个序列段的起始点和各个序列段的表示符号对时间序列数据进行近似描述。该算法的基本思想如下：

1)取出时间序列 X 中的全局最大值 $x_{\max}$ 和全局最小值 $x_{\min}$，并利用式(1)进行归一化操作；再对归一化后的时间序列数据进行标准化，即将其标准化为正态分布均值为 0、标准差为 1 的标准时间序列 X'。

$$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{1}$$

2)对 X'进行等长分段，并对 X'进行近似表示。

3)提出始末距离的概念，并结合符号距离定义新的距离度量公式。

3.1 近似表示

给定 n 维时间序列 X，得到标准时间序列 X'，将标准时间序列 X'等分为 ω 个子序列段(若不能等分，则最后一个子序列段取 $n-\left\lceil\frac{n}{\omega}\right\rceil*(\omega-1)$个维度)。去掉各个子序列段的起点和终点后，按照式(2)计算各段的近似均值$\overline{X_i}$：

$$\overline{X_i}=\frac{1}{\frac{n}{\omega}-2}\sum_{\frac{(i-1)n}{\omega}+2}^{\frac{in}{\omega}-1}x_i',\ 1\leqslant i\leqslant\omega \tag{2}$$

根据原符号聚合近似表示方法(SAX)中指定的字母表大小 α 和分割点 β_i 对各个序列段的近似均值$\overline{X_i}$进行符号化赋值，得到各个序列段的表示符号$\hat{X}_i$($1\leqslant i\leqslant\omega$)。

取各个序列段的表示符号$\hat{X}_i$以及各个序列段起点和终点的索引值组成一个低维的特征向量，用来近似表示原时间序列 X，表示方式如式(3)所示：

$$\widetilde{X}=((s_1,\hat{X}_1),e_1),(s_2,\hat{X}_2,e_2),\cdots,(s_\omega,\hat{X}_\omega,e_\omega)) \tag{3}$$

其中，s_i 表示第 i 段起点的索引值，e_i 表示第 i 段终点的索引值。由于第 $i(i\neq\omega)$段的终点的索引值是第 $i+1$ 段起点的索引值的前一位，因此在近似表示原时间序列时，除第 ω 个序列段需存储终点的索引值外，其他序列段都只需存储起点的索引值，如式(4)所示：

$$\widetilde{X}=(1,\hat{X}_1,\frac{n}{\omega}+1,\hat{X}_2,\frac{2n}{\omega}+1,\hat{X}_3,\cdots,\frac{(\omega-1)n}{\omega}+1,\hat{X}_\omega,n) \tag{4}$$

则相比于原表示方法，时间序列 X 的近似表示方法也只增加了 $\omega+1$ 维。对于庞大的时间序列数据而言，增加 $\omega+1$ 维的存储空间是可以接受的。

3.2 相似度度量

时间序列数据有明显的形态趋势，即随着时间序列出现上升、下降或者水平的状态，原符号聚合近似表示方法 SAX 中提出的距离度量仅通过符号距离对两个时间序列进行距离度量，忽略了时间序列各序列段的形态距离。本文方法针对第 3 节首段提出的形态特征(各个序列段的起点和终点)，基于欧氏距离的思想定义了始末距离公式，以计算两序列段间的形态距离，并结合由本表示方法得到的各个序列段的表示符号，给出新的距离度量公式。

定义 1(始末距离) 对于两个长度相同的序列段 P 和

Q,它们之间的始末距离由式(5)计算:

$$smd(P,Q)=\sqrt{(P_s-Q_s)^2+(P_e-Q_e)^2} \tag{5}$$

其中,P_s 和 Q_s 表示所述子序列段 P 和 Q 的起点值,P_e 和 Q_e 表示所述子序列段 P 和 Q 的终点值。

图 1 中给出的 4 条序列段的均值一致,使用式(5)来计算始末距离,可以得到序列段 a 和序列段 c 之间的始末距离最短,序列段 b 和序列段 d 之间的始末距离最短,这与我们的常规判断是一致的。

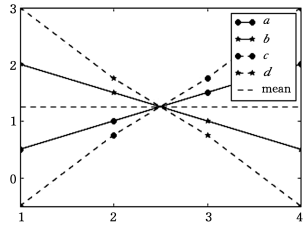

图 1　4 条序列段的趋势

Fig. 1　Trend of 4 sequence segments

序列段 P 和 Q 之间的符号距离可以通过式(6)进行计算:

$$dist(P,Q)=\begin{cases}0, & \text{if } |\hat{P}-\hat{Q}|\leqslant 1\\ \beta_{\max(\hat{P},\hat{Q})-1}-\beta_{\min(\hat{P},\hat{Q})}, & \text{otherwise}\end{cases} \tag{6}$$

其中,$\hat{P}$ 和 $\hat{Q}$ 为所述各段的表示符号,β_i 表示高斯曲线第 i 个分割点。

假设存在两个时间序列 X 和 Y,本文基于始末距离和符号距离提出了如下时间序列的距离度量公式:

$$D(X,Y)=\sqrt{\frac{n}{\omega}}\sqrt{\sum_{i=1}^{\omega}(dist(\hat{X}_i,\hat{Y}_i)^2)+\frac{\omega}{n}(\sum_{i=1}^{\omega}smd(X_i,Y_i)^2)} \tag{7}$$

该相似度度量公式克服了纯符号距离的弊端,也结合了形态特征,能够更准确、更详细地测量时间序列间的相似性。

4　实验及理论分析

4.1　下界性理论证明

给定两条时间序列数据,X 由 n 个数据点$(1,x_1),(2,x_2),\cdots,(n,x_n)$组成,$Y$ 也由 n 个数据点$(1,y_1),(2,y_2),\cdots,(n,y_n)$组成,则这两条时间序列之间的欧氏距离为:

$$D_1(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{8}$$

由原符号聚合近似表示方法(SAX)可知:

$$D_2(X,Y)=\sqrt{\frac{n}{\omega}}\sqrt{\sum_{i=0}^{\omega}(\overline{X_i}-\overline{Y_i})^2}\leqslant D_1(X,Y) \tag{9}$$

现需要证明式(7)≤式(8),首先对式(7)进行如下转化,将符号距离转化为近似均值距离,得式(10):

$$D_3(X,Y)=\sqrt{\frac{n}{\omega}}\sqrt{\sum_{i=1}^{\omega}((\overline{x_{(i-1)\frac{n}{\omega}+2}\sim x_{i\frac{n}{\omega}-1}}-\overline{y_{(i-1)\frac{n}{\omega}+2}\sim y_{i\frac{n}{\omega}-1}})^2)+\frac{\omega}{n}(\sum_{i=1}^{\omega}smd(X_i,Y_i)^2)} \tag{10}$$

其中,$\overline{x_{(i-1)\frac{n}{\omega}+2}\sim x_{i\frac{n}{\omega}-1}}$,$\overline{y_{(i-1)\frac{n}{\omega}+2}\sim y_{i\frac{n}{\omega}-1}}$分别表示 X 和 Y 第 i 段除去起点和终点的近似均值。

则只要证明:

$$D_3(X,Y)\leqslant D_1(X,Y) \tag{11}$$

根据式(10)和式(8),可以得到:

$$D_3(X,Y)^2=\frac{n}{\omega}(\overline{x_{(i-1)\frac{n}{\omega}+2}\sim x_{i\frac{n}{\omega}-1}}-\overline{y_{(i-1)\frac{n}{\omega}+2}\sim y_{i\frac{n}{\omega}-1}})^2+\sum_{i=1}^{\omega}[(x_{(i-1)\frac{n}{\omega}+1}-y_{(i-1)\frac{n}{\omega}+1})^2+(x_{i\frac{n}{\omega}}-y_{i\frac{n}{\omega}})^2] \tag{12}$$

$$D_1(X,Y)^2=(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2$$

$$D_1(X,Y)^2=\sum_{i=1}^{\omega}[(x_{(i-1)\frac{n}{\omega}+1}-y_{(i-1)\frac{n}{\omega}+1})^2+(x_{i\frac{n}{\omega}}-y_{i\frac{n}{\omega}})^2]+\sum_{i=1}^{\omega}\sum_{j=2}^{\frac{n}{\omega}-1}((x_{(i-1)\frac{n}{\omega}+j}-y_{(i-1)\frac{n}{\omega}+j})^2) \tag{13}$$

由式(12)和式(13)可得,要证式(11)成立,只需要证:

$$\frac{n}{\omega}(\overline{x_{(i-1)\frac{n}{\omega}+2}\sim x_{i\frac{n}{\omega}-1}}-\overline{y_{(i-1)\frac{n}{\omega}+2}\sim y_{i\frac{n}{\omega}-1}})^2\leqslant\sum_{i=1}^{\omega}\sum_{j=2}^{\frac{n}{\omega}-1}((x_{(i-1)\frac{n}{\omega}+j}-y_{(i-1)\frac{n}{\omega}+j})^2) \tag{14}$$

由已知条件式(9)可得:

$$\begin{aligned}&\frac{n}{\omega}(\overline{x_{(i-1)\frac{n}{\omega}+2}\sim x_{i\frac{n}{\omega}-1}}-\overline{y_{(i-1)\frac{n}{\omega}+2}\sim y_{i\frac{n}{\omega}-1}})^2\\&\leqslant(x_2-y_2)^2+\cdots+(x_{\frac{n}{\omega}-1}-y_{\frac{n}{\omega}-1})^2+(x_{\frac{n}{\omega}+2}-y_{\frac{n}{\omega}+2})^2+\cdots+(x_{2\frac{n}{\omega}-1}-y_{2\frac{n}{\omega}-1})^2+\cdots+(x_{(\omega-1)\frac{n}{\omega}+2}-y_{(\omega-1)\frac{n}{\omega}+2})^2+\cdots+(x_{n-1}-y_{n-1})^2\\&=\sum_{i=1}^{\omega}\sum_{j=2}^{\frac{n}{\omega}-1}((x_{(i-1)\frac{n}{\omega}+j}-y_{(i-1)\frac{n}{\omega}+j})^2)\end{aligned} \tag{15}$$

因为不等式(14)成立,所以不等式(11)成立,即式(7)≤式(8),因此基于始末距离的符合聚合近似表示方法中的距离度量公式满足下界定理。

4.2　实验环境和实验数据

实验环境:编译软件/python2.7.12,操作系统/Windows 7,CPU/Intel(R) Core(TM) i5-2450M,主频 2.50 GHz,内存 8 GB,硬盘容量 500 GB。

实验数据:从 UCR 数据集中抽取了 20 组公开数据集,通过 UCR 数据集来分析各类时间序列相关算法的性能,因此将这 20 组数据用于本方法的实验也是有意义的。数据描述如表 2 所列。

表 2　20 组数据集的详细信息

Table 2　Detailed information of twenty data sets

数据集名称	类别数	训练集	测试集	维度
Syn_Control	6	300	300	60
Olive Oil	4	30	30	570
Gun_Point	2	50	150	150
CBF	3	30	900	128
FaceAll	14	560	1690	131
OSU Leaf	6	200	242	427
Swedish_Leaf	15	500	625	128
50words	50	450	455	270
Trace	4	100	100	275
Two Patterns	4	1000	4000	128
Wafer	2	1000	6174	152
FaceFour	4	24	88	350
Lighting_2	2	60	61	637
Lighting_7	7	70	73	319
ECG	2	100	100	96
Adiac	37	390	391	176
Yoga	2	300	3000	426
Beef	5	30	30	470
Fish	7	175	175	463
Coffee	2	28	28	286

这 20 组数据集都被分为训练集和测试集，且数据集的类别数从 2 类跨度到 50 类，实验范围比较广泛，保证了实验的全面性和代表性。

4.3 分类准确率

本文采用 1 最近邻分类算法(1NN)来比较不同表示方法中相似度度量公式的分类准确率。首先采用 1NN 算法对测试集中的数据与训练集中的数据进行距离计算，然后选取距离最小的样本的类别对测试集进行预测分类。如果测试样本的预测分类与真实类别一致，则分类准确；否则分类错误。分类准确率即为测试集中分类准确的样本数与测试集的总样本数之比。1NN 算法能够直接决定距离度量的有效性。本文中用于对比的距离变量方法包括欧氏距离、SAX 距离、SAX_TD 距离。欧氏距离是目前时间序列中应用最为广泛的一种距离，SAX_TD 距离是在 SAX 距离的基础上提出的一种改进，本文提出的 SAX_SM 距离是在 SAX_TD 距离的基础上进行的改进，因此本文比较这 4 种距离度量。

不同的字母表大小 α 和不同的分段数目 ω 能使算法的效果达到不同的分类准确率。本文为了使 4.1 节中的 20 组时间序列数据集得到最好的分类准确率，将字母表大小 α 取为 10；限制分段数目 ω 从 2 到 $n/2$ 取值，且每次取值都是前一次取值的 2 倍。如果采用不同的参数取得了相同的分类准确率，则选取 ω 较小的参数。

表 3 各种方法在不同数据集中的分类准确率

Table 3 Classification accuracy of each method in different data sets

数据集	Euclid	SAX	SAX_TD	SAX_SM
Syn_Control	0.880	**0.977**	0.973	**0.977**
Olive Oil	**0.867**	0.167	**0.867**	**0.867**
Gun_Point	0.910	0.800	0.907	**0.953**
CBF	0.852	0.893	0.908	**0.910**
FaceAll	**0.714**	0.677	0.708	0.670
OSU Leaf	0.521	**0.541**	0.537	**0.541**
Swedish Leaf	0.788	0.515	**0.806**	0.789
50words	0.631	**0.659**	0.651	**0.659**
Trace	**0.760**	0.62	0.750	0.730
Two Patterns	0.907	0.919	0.929	**0.931**
Wafer	0.995	**0.996**	0.995	**0.996**
FaceFour	0.784	**0.830**	**0.830**	0.818
Lighting_2	0.754	**0.787**	0.771	**0.787**
Lighting_7	0.575	0.603	**0.685**	0.644
ECG	0.880	0.880	**0.900**	0.880
Adiac	0.611	0.105	0.604	**0.634**
Yoga	**0.830**	0.802	0.821	0.827
Beef	0.667	0.500	0.667	**0.700**
Fish	0.783	0.486	**0.794**	**0.794**
Coffee	**1.000**	0.536	**1.000**	**1.000**

由表 3 中的结果(最大的分类准确率已加黑)可知，在 20 组数据集中，与欧氏距离相比，SAX_SM 分类在 17 组数据集上的准确率达到甚至超过欧氏距离，其中在 14 组数据集上超过欧氏距离；在 Syn_Control 数据集上，分类准确率提升最高达到 9.7%；在 Wafer 数据集上，分类准确率提升最小，仅 0.1%。与 SAX 相比，SAX_SM 的分类准确率在 18 组数据集上达到甚至超过 SAX，其中在 12 组数据集上超过 SAX；在 Olive Oil 数据集上，分类准确率提升最大，达 70%；在 Two Patterns 数据集上，分类准确率提升最小，仅 1.2%。与 SAX_TD 相比，SAX_SM 的分类准确率在 14 组数据集上达到甚至超过 SAX_TD，其中在 11 组数据集上超过 SAX_TD；在 Gun_Point 数据集上，分类准确率提升最大，达 4.6%；在 Wafer 数据集上，分类准确率提升最小，仅 0.1%。因此，SAX_SM 的分类效果较欧氏距离、SAX 和 SAX_TD 更优。

4.4 算法的运行时间

本节对 SAX，SAX_TD，SAX_SM 的运行时间进行比较。选择 5 组不同维度的时间序列数据集来测试 3 种表示方法的运行时间，这 5 组时间序列数据集分别为 Syn_Control，ECG，Gun_Point，Trace 和 Olive Oil，它们的维度 n 分别为 60，96，150，275 和 570，则它们的最大分段数目 ω 取值分别为 16，32，64，128 和 256，选择符号表大小 α 为 10 时来进行实验比较。实验结果如图 2－图 6 所示。

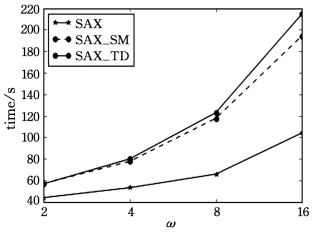

图 2 Syn_Control 数据集

Fig. 2 Syn_Control data set

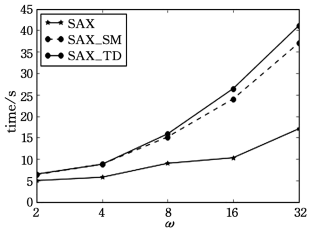

图 3 ECG 数据集

Fig. 3 ECG data set

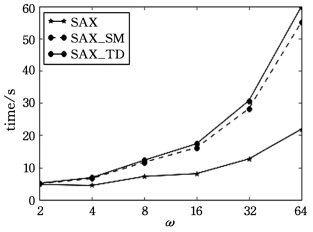

图 4 Gun_Point 数据集

Fig. 4 Gun_Point data set

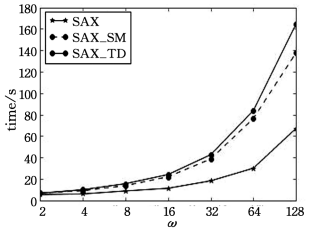

图 5 Trace 数据集

Fig. 5 Trace data set

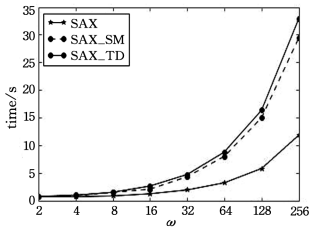

图 6 Olive Oil 数据集

Fig. 6 Olive Oil data set

从图 2－图 6 可以看出，随着分段数目的增加，3 种方法的运行时间都在不断增加。SAX_SM 方法的运行时间虽然比 SAX 方法的长，但比 SAX_TD 的短，且随着分段数目的增加，较 SAX_TD 的减少程度更加明显。相对于 SAX_TD 方法，SAX_SM 方法提高了效率。

结束语 基于原符号聚合近似表示方法无法解决当时间序列数据集中各个序列段表示符号一致的情况，本文提出了

基于始末距离的时间序列符号聚合近似表示方法,在使用表示符号的同时,加上了始末特征,综合符号信息和始末信息对时间序列数据集进行近似描述,并给出了新的距离度量。SAX_SM 方法的存储空间比 SAX 增加了 $\omega+1$ 维,但是在时间序列数据维度较多的情况下,增加 $\omega+1$ 维的存储空间也是可以接受的,且增加的维度与 SAX_TD 方法的一致。经理论证明,该表示方法提出的距离度量满足下界定理。

本文在 20 组 UCR 时间序列数据集上进行了实验。实验结果表明,本文提出的 SAX_SM 方法在 13 个数据集中获得了最高的分类准确率(包含并列最大的),SAX_TD 在 7 个数据集中获得了最高的分类准确率(包含并列最大的),而 SAX 只在 6 个数据集中获得了最高的分类准确率(包含并列最大的),且 SAX_SM 方法的运行时间介于原符号聚合表示方法和改进的符号聚合近似表示方法之间,从而说明 SAX_SM 具有显著优于 SAX 和 SAX_TD 的分类性能。

因为本文提出的初步实验结果主要集中在距离度量上,所以未来计划将该表示方法有效地应用于其他数据挖掘任务中,如聚类、分类等。

参 考 文 献

[1] CRYER J D,CHAN K S,时间序列分析及应用:R 语言[M]. 潘红宇,译. 北京:机械工业出版社,2011:25-29.

[2] HAN J,PEI J,KAMBER M. Data mining:concepts and techniques[M]. Amsterdam:Elsevier,2011:20-23.

[3] FU T. A review on time series data mining[J]. Engineering Applications of Artificial Intelligence,2011,24(1):168-181.

[4] LI H L. Research on Feature Representation and Similarity Meaure Methods in Time Series Data Mining[D]. Dalian:Dalian University of Technology,2012. (in Chinese)
李海林. 时间序列数据挖掘中的特征表示与相似性度量方法研究[D]. 大连:大连理工大学,2012.

[5] ESLING P,AGON C. Time-series data mining[J]. ACM Computing Surveys (CSUR),2012,45(1):12.

[6] LI H L,GUO C H. Survey of feature representations and similarity measurements in time series data mining[J]. Application Research of Computers,2013,30(5):1285-1291. (in Chinese)
李海林,郭崇慧. 时间序列数据挖掘中特征表示与相似性度量研究综述[J]. 计算机应用研究,2013,30(5):1285-1291.

[7] YUAN J D,WANG Z H. Review of Time Series Representation and ClassificationTechniques [J]. Computer Science, 2015, 42(3):1-7. (in Chinese)
原继东,王志海. 时间序列的表示与分类算法综述[J]. 计算机科学,2015,42(3):1-7.

[8] AGRAWAL R,FALOUTSOS C,SWAMI A. Efficient similarity search in sequence databases[C]// International Conference on Foundations of Data Organization and Algorithms. 1993:69-84.

[9] RATANAMAHATANA C,KEOGH E,BAGNALL A J,et al. A novel bit level time series representation with implication of similarity search and clustering[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg,2005:771-777.

[10] 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.

[11] AZZOUZI M,NABNEY I T. Analysing time series structure with Hidden Markov Models[C]// Neural Networks for Signal Processing VIII,1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop. IEEE,1998:402-408.

[12] KALPAKIS K,GADA D,PUTTAGUNTA V. Distance measures for effective clustering of ARIMA time-series[C]// Proceedings IEEE International Conference on Data Mining,2001(ICDM 2001). IEEE,2001:273-280.

[13] KEOGH E,CHAKRABARTI K,PAZZANI M,et al. Dimensionality reduction for fast similarity search in large time series databases[J]. Knowledge and Information Systems,2001,3(3):263-286.

[14] ZHU Y. High performance data mining in time series:techniques and case studies[D]. New York:New York University,2004.

[15] LKHAGVA B,SUZUKI Y,KAWAGOE K. New time series data representation ESAX for financial applications[C]// 22nd International Conference on Data Engineering Workshops. IEEE,2006:115-115.

[16] ZHONG Q L,CAI Z X. The Symbolic Algorithm for Time Series Data Based on Statistic Feature[J]. Chinese Journal of Computers,2008,31(10):1857-1864. (in Chinese)
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.

[17] ESMAEL B,ARNAOUT A,FRUHWIRTH R,et al. Multivariate time series classification by combining trend-based and value based approximations[M]// Computational Science and Its Applications-ICCSA 2012. Springer Berlin Heibelberg,2012:392-403.

[18] SUN Y,LI J,LIU J,et al. An improvement of symbolic aggregate approximation distance measure for time series[J]. Neurocomputing,2014,138(11):189-198.

[19] ZHENG X,SHENG L H,CUI X Y. A Piecewise Aggregation Approximation of Time Series Based on Wavelet Entropy [J]. Computer Simulation,2015,32(1):411-415. (in Chinese)
郑旭,盛立辉,崔宵语. 基于小波熵的时间序列分段聚合近似表示[J]. 计算机仿真,2015,32(1):411-415.

[20] WANG Y,AN Y. Composite similarity measure algorithm[C]// 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery(ICNC-FSKD). IEEE,2016:1254-1258.

[21] LARSEN R J,MARX M L. An introduction to mathematical statistics and its applications[M]. Prentice-Hall Englewood,1986:470-481.