http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0332

基于趋势符号聚合近似的卫星时序数据分类方法

阮辉[1]，刘雷[2]，胡晓光[1,*]

(1. 北京航空航天大学 自动化科学与电气工程学院，北京 100083； 2. 北京机电工程研究所，北京 100074)

摘 要： 作为在时间序列数据挖掘中广泛使用的主要符号化表示方法，符号聚合近似(SAX)使用段的平均值作为符号表示，由于无法区分具有不同趋势但具有相同平均值符号的不同时间序列，某些情况下可能会导致错误的分类。提出了一种改进的符号表示——趋势符号聚合近似(TrSAX)，集成SAX与最小二乘法，用以描述时间序列的均值和斜率，并由此构建出BOTS分类器。此外，对卫星的模拟量遥测时序数据中的角度序列、转速序列、电流序列进行分析，并从UCR公开数据集中筛选出与3种序列类似的3个数据集进行分类实验验证。与应用了SAX和2个改进的SAX、经典的欧氏距离(ED)、动态时间规整(DTW)的1-NN分类方法进行对比，结果表明：提出的BOTS分类方法的分类错误率明显低于其他5种分类方法。

关 键 词： 卫星遥测数据；时间序列；符号化表示；时间序列；异常检测

中图分类号： V574；TP311

文献标志码： A **文章编号：** 1001-5965(2021)02-0333-09

时间序列是按照时间排序的一组随机变量，其通常是在相等间隔的时间段内依照给定的采样率对某种潜在过程进行观测的结果[1]。在卫星的测控管理过程中，会产生大量的遥测数据，它们以时间序列的形式存储在数据库中。而运行状态监测系统传感器产生的监测数据通过遥测系统传输至地面控制中心，此类数据是地面判断在轨卫星运行和健康状态的唯一依据[2]。同时，这些海量的时间序列蕴含可用于卫星故障诊断的规律和知识。通过数据挖掘，可以提取卫星各器件的信息，发现异常、关联、模式、趋势等知识。有效掌握和利用这些信息规律对卫星异常检测和关联分析、故障诊断、监测预警，对卫星测管管理与决策活动，如改进卫星设计、提高测试及监测自动化等工作具有特别重要的意义[3-7]。

此外，在金融[8]、天气观察[9]、生物医学测量[10]领域同样大量产生此类型的数据。因此，在过去几十年中，时间序列数据挖掘的研究在开发新算法和改进可用算法以满足当前需求方面一直非常活跃。

由于原始时间序列具有高维、高音量和大量噪声的特征，计算机要对原始时间序列进行分类面临许多挑战。

现有的分类算法可以分为2类，即基于形状的分类算法和基于结构的分类算法。

基于形状的分类算法以最近邻(One-Nearest-Neighbor，1-NN)分类器作为基础，时序数据的表示方法则为欧氏距离(Euclidean Distance，ED)[11]和动态时间规整(Dynamic Time Warping，DTW)[12]。

在短时间序列数据集上，1-NN DTW已被证明是具有高度代表性和竞争力的分类器。基于形状的技术的不足之处在于：当使用噪声或包含特征子结构的长数据对数据进行分类时，它们显示的性能很差。

收稿日期：2020-07-12；录用日期：2020-08-07；网络出版时间：2020-08-18 10:02
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20200817.1820.002.html
基金项目：国家自然科学基金(51807003)；国防基础科研计划(JKCY2016204A102)

* 通信作者. E-mail：xiaoguang@buaa.edu.cn

引用格式：阮辉，刘雷，胡晓光. 基于趋势符号聚合近似的卫星时序数据分类方法[J]. 北京航空航天大学学报，2021，47(2)：333-341. RUAN H，LIU L，HU X G. Satellite time series data classification method based on trend symbolic aggregation approximation[J]. Journal of Beijing University of Aeronautics and Astronautics，2021，47(2)：333-341(in Chinese).

基于结构的技术有2个步骤。首先,这些方法包括离散傅里叶变换(Discrete Fourier Transform, DFT)[13],可索引分段线性逼近[14]用于提取特征向量;然后,采用经典数据挖掘算法(如决策树),支持向量机被用来做分类任务。

目前,符号化方法也被广泛使用。因为它具有简单、可读性、高效率之外,还可以使用其他领域的算法,如信息检索或文本处理等。其中,Lin等[15]提出的符号聚合近似(Symbolic Aggregate Approximation, SAX)方法是时下流行的符号化表示方法。SAX基于分段总体逼近(Piecewise Aggregate Approximation, PAA)方法[16],并假定PAA值遵循高斯分布。SAX根据高斯曲线下的等大小区域离散化PAA值,从而产生断点,后采用计算开销小的低边界计算准确的距离。

模式袋模型(Bag-of-Patterns, BOP)[17]提取子结构作为时间序列的高级特征,将这些子结构转化为SAX,并采用ED作为度量距离进行相似性度量的相关应用,如分类、聚类等。BOP方法不考虑子结构之间的顺序,只是将时间序列看成是一些SAX字符串出现概率的集合,每个SAX字符串是相互独立的,类似于直方图的统计表示。

但SAX表示的质量取决于PAA系数,即时间序列分割的段数,用于量化的符号数量(字母大小数)和高斯假设。同时,SAX中的符号是根据每个分段的均值绘制的,不能表示分段的趋势。

为此,一些文献尝试解决SAX的这些问题。Pham等[18]通过引入对时间序列截的长度和字母大小起作用的自适应断点矢量来缓解高斯假设。但是,通过使用聚类方法引入预处理阶段,就失去了SAX的简单性。为描述SAX的趋势信息,文献[19]中提出的扩展符号聚合近似(Extended SAX, ESAX)将PAA段的符号最小值和最大值与相关的SAX符号以及它们的出现顺序相关联。这定义了一个抽象形状,可以在时间序列检索中提供更好的结果。但是,ESAX表示的大小是SAX表示的大小的3倍。从效率的角度来看,笔者没有将文献[19]的方法与相同大小的SAX表示进行比较。文献[20]提出的基于趋势的符号聚合逼近(Trend-based Symbolic Aggregate approXimation, TSAX),在每个分段中添加了2个趋势指标,TSAX可以区分2个不同趋势之间的差异,但仍不能反映同一趋势中的差异。另外,TSAX表示的大小与ESAX相同,是SAX的3倍。

提高分类算法准确率、减少运算时间是数据符号化表示的主要目标。本文提出一种改进的符号表示——趋势符号聚合近似(Trend Symbolic Aggregate Approximation, TrSAX),集成SAX与最小二乘法,用以描述时间序列的均值和斜率,在不增加运算时间的情况下,进一步提高了分类的准确率。

1 卫星时序数据

1.1 卫星遥测数据来源

根据任务的不同,一颗卫星所含分系统略有差别,但通常包括热控分系统、姿轨控分系统、星务管理分系统、电源分系统、推进分系统、结构分系统、测控分系统、总体电路分系统等。

为监测各分系统的工作状态,设计时在各分系统中设置了大量的测试传感器,采集测试信息传递到地面,形成卫星遥测数据。

1.2 卫星遥测数据特点

卫星的测试数据一般分为数字量和模拟量。数字量可细分为独立数字量、关联数字量和状态数字量。模拟量可细分为恒定模拟量、区间模拟量和趋势变化模拟量。其中,数字量主要为指令、计数和状态等,模拟量主要为电流、电压、角度、温度、压力等。此外,由于卫星是按照轨道周期运行,卫星遥测数据存在周期性。

综上,不同卫星不同分系统的具体遥测数据需要在进行充分分析、数据预处理之后,进行分类、聚类、异常检测等相关研究。

1.3 真实卫星遥测数据分析

文献[21]中对“风云三号”卫星遥测数据进行了分析,其中含133个测试参量,时间跨度为500天。并对3种具有周期特性的测试序列(电流、角度、转速)为对象进行了详细研究。

由于卫星遥测数据具有明显的周期性,通过对遥测数据的每个周期进行分析,可知卫星在该周期之内的运行状态是否正常。

首先,需要对遥测数据进行分段,具体根据卫星工作方式进行分段,根据幅角测试参量的测试值为0~360不断循环,具有周期性,且周期分隔点明确。同时,幅角测试值能够反映出卫星轨道的运行位置,因此,将测试值由360跳变到0的点作为幅角突变点进行分段。对于因数据缺失而不完整的序列,在实验中进行剔除,确保各分段子序列完整。

如图1~图3所示,以角度测试时间序列的幅角突变点为标识进行分段,将每段序列进行叠加绘图。样本数据样本选用的是3种遥测数据测

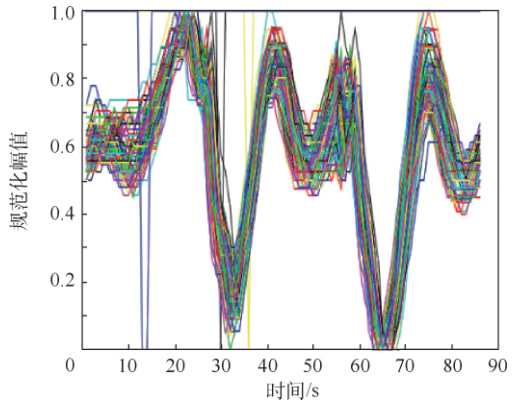

图 1　卫星某角度测试时间序列分段叠加结果

Fig. 1　Satellite angle test time series segmentation merging results

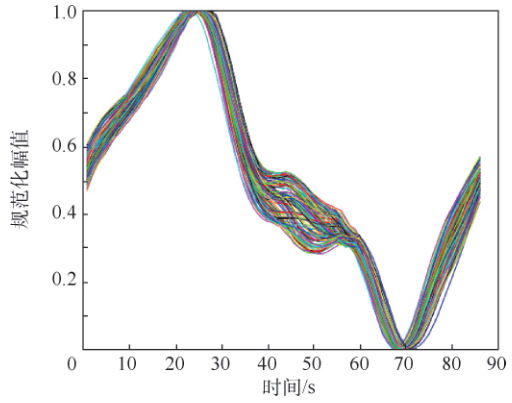

图 2　卫星某转速测试时间序列分段叠加结果

Fig. 2　Satellite rotation speed test time series segmentation merging results

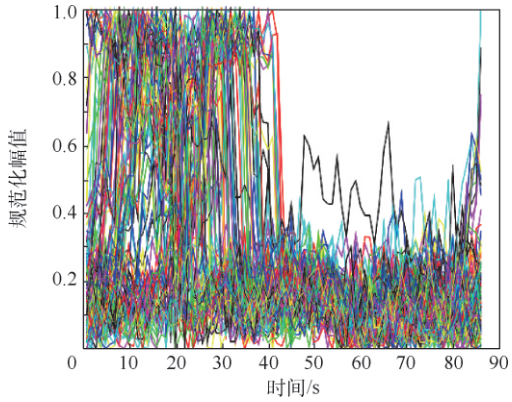

图 3　卫星某电流测试时间序列分段叠加结果

Fig. 3　Satellite electric current test time series segmentation merging results

试序列的前 4 000 个有效样本点,每个周期的数据点为 86,样本数据包含 47 个周期。图中显示各个分段子序列之间的耦合度高,分段合理。

图中分段子序列具有以下特点:

1) 各子序列整体变化趋势相同,局部有细小差别,波形不完全重合。

2) 测试子序列波动趋势不同,有的较为平滑如图 2 所示,有的较为激烈如图 3 所示。

2　相关定义

2.1　时间序列

时间序列数据[15]是指将使用同一统计指标的数值按照时间先后顺序排列而成的数列:

$$T = (t_1, t_2, \cdots, t_n) \tag{1}$$

式中: 时间序列 T 的长度为 n。

2.2　截

使用滑动窗口函数[17],将时间序列 $T = (t_1, t_2, \cdots, t_n)$ 分为固定长度为 ω 的多个截 $S_{i;\omega} = (t_i, t_{i+1}, \cdots, t_{i+\omega-1})$。

$$\text{Segment}(T, \omega) = \{S_{1;\omega}, S_{2;\omega}, \cdots, S_{n-\omega+1;\omega}\} \tag{2}$$

式中: 时间序列 T 分成 $n-\omega+1$ 个截。

$$\begin{cases} S_{1;\omega} = (t_1, t_2, \cdots, t_\omega) \\ S_{2;\omega} = (t_2, t_3, \cdots, t_{\omega+1}) \\ \vdots \\ S_{n-\omega+1;\omega} = (t_{n-\omega+1}, t_{n-\omega+2}, \cdots, t_n) \end{cases} \tag{3}$$

2.3　段

截 $S_{i;\omega} = (t_i, t_{i+1}, \cdots, t_{i+\omega-1})$ 被分成 m 个等长的段,表示为集合 $\{s_j, j = 1, 2, \cdots, m\}$, $s_j = \{t_{(j-1)\times L+l}, l = 1, 2, \cdots, L\}$, $L = \text{floor}(\omega/m)$, floor 函数为向下取整函数。

图 4 为显示了时间序列划分的示意图。长度为 512 的时间序列样本被分为 449 个截,每个截被等分为 4 段。

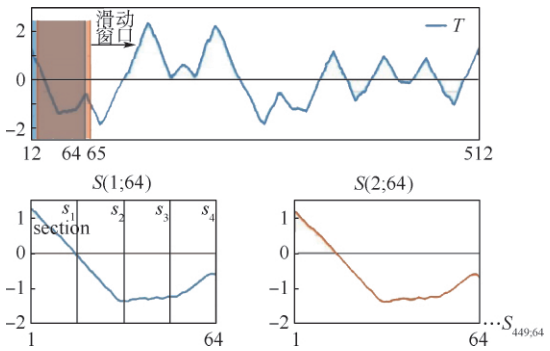

图 4　时间序列划分示意图

Fig. 4　Schematic diagram of time series division

3　符号化表示

3.1　段平均值

段 $s_j = \{t_{(j-1)\times L+l}, l = 1, 2, \cdots, L\}$ 的平均值 $\bar{s}_j$ 计算如下:

$$\bar{s}_j = \frac{1}{L}\sum_{i=(j-1)\times L+1}^{j\times L} t_i \tag{4}$$

表 1 列出了字母数为 3 ~ 9 的断点,断点字母符号的范围为3 ~ 5。使用 $\varphi - 1$ 个断点将分布空

间划分为 φ 个等概率区域。$N(0,1)$ 高斯曲线下的 $B=\beta_1,\beta_2,\cdots,\beta_i$ 的断点间面积等于 $1/\varphi$。

使用这些断点，将每个部分的平均值 $\bar{s}_j$ 绘制为对应于其所在区域的小写字母，如图 5 所示。

表 1 字母数为 3～9 的断点查找表

Table 1 Look up table from breakpoints with alphabet sizes from 3 to 5

β_i	3	4	5
β_1	−0.43	−0.67	−0.84
β_2	0.43	0	−0.25
β_3		0.67	0.25
β_4			0.84

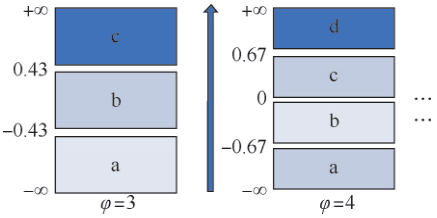

图 5 $\varphi=3$ 和 $\varphi=4$ 时，各平均值和小写字母之间的对应关系

Fig. 5 Corresponding relationship between average values and lower-case letters for $\varphi=3$ and $\varphi=4$

3.2 段斜率值

SAX 忽略了时间序列段中对于分析时间序列分类和相似性至关重要的趋势信息这一重要特征。为了描述趋势信息，本文使用最小二乘法来计算每个时间序列的斜率值。

$$k=\frac{L\sum_{j=1}^{L}(j\times t_j)-\sum_{j=1}^{L}j\sum_{j=1}^{L}t_j}{L\sum_{j=1}^{L}t_j^2-2\sum_{j=1}^{L}t_j} \tag{5}$$

式中：$L=\mathrm{floor}(\omega/m)$，为一个截中段的个数；$j$ 表示第 j 个段。

采用 0°和 ±45°三个角度的斜率作为角度间隔值将角度空间(−90°,90°)划分为 5 个非重叠间隔，如图 6 所示，将计算出的斜率用大写字母表示，该字母对应于它们的驻留区域。通过这种方法，将角度空间(−90°,90°)转换为 5 个大写字母，代表 5 种情况：速降(A)、缓降(B)、水平(C)、缓增(D)和速增(E)。

每段用 2 个符号表示，大写字母表示斜率值，小写字母表示平均值。将计算出的大写字母和小写字母组合起来代表每个时间序列截。这些为每个时间截组成的大小字母串，称为趋势符号聚合近似(TrSAX)表示。图 7 为长度 64 的时间序列截的 TrSAX 字母串，对于 $\varphi=4$，截的 TrSAX 字为 AcAaDaEa。

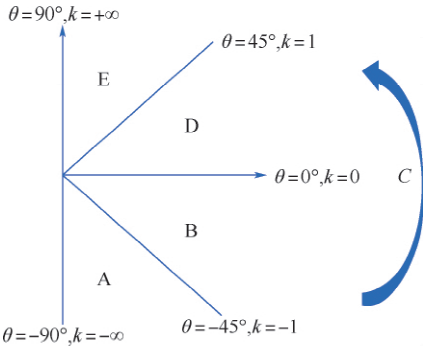

图 6 角度间隔和字母的对映关系

Fig. 6 Angle interval and corresponding letters

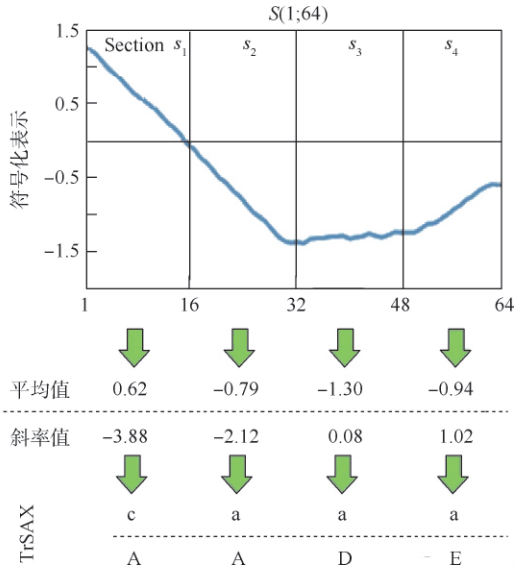

图 7 趋势符号聚合近似(TrSAX)表示示意图

Fig. 7 Schematic diagram of trend symbolic aggregate approximation (TrSAX) representation

4 TrSAX 词袋

与 BOP 方法类似，本文提取时间序列中的截作为子序列，并将其转换为 TrSAX 词，则一条时间序列可转为一个 TrSAX 词袋(Bag-of-TrSAX，BOTS)。

4.1 BOTS 参数设定

BOTS 有 4 个参数，分别为截长度 ω 和 3 个用于表示截的 TrSAX 参数，即字母符号数 φ、该截中段的数量 m 和大写字母数 γ。

4.2 BOTS 工作流程

1) 使用 4.1 节的 TrSAX 参数，为数据集建立一个大小为 $s=(\gamma\varphi)^m$ 的 TrSAX“词汇表”。

2) 使用滑动窗口函数，将长度为 n 的原始时间序列 T 转换为 $n-\omega+1$ 个固定大小的截。

3) 将这些时间序列截进行(0,1)标准化。

4) 将时间截转换为 TrSAX 字。

5) 时间序列 T 转换成为一个 TrSAX 字袋。

如果时间序列没有噪声或尖峰,是平滑的,则许多连续的截将转换为相同的 TrSAX 字。若对所有这些出现的 TrSAX 单词进行计数,则计数过多。

为此,为了防止这种情况,本文使用数据规约[17],仅统计重复出现的 TrSAX 词中第一个,并忽略所有重复项,直到新词出现。数据规约如下所示,首次出现的 TrSAX 词标粗。

假设获得如下 TrSAX 字符串 S:

S = **DbDbDb** DbDbDb **DbDbDc** DbDbDc DbDbDc **DbDcBc** DbDcBc **DcBcBc** DcBcBc **BcBcBb** BcBcBb BcBcBb **BcBbBb** BcBbBb **BbBbBb** BbBbBb **BbBbDb** BbBbDb **BbDbDb** BbDbDb **DbDbDb** DbDbDb…

数据规约之后,字符串 S 迭代为 S':

S' = **DbDbDb DbDbDc DbDcBc DcBcBc BcBcBb BcBbBb BbBbBb BbBbDb DbDbDb**…

然后,构造直方图 H 以计算出现在字符串 S' 中的 TrSAX 单词的数量。

H: DbDbDb = 2, DbDbDc = 1, DbDcBc = 1, DcBcBc = 1, BcBcBb = 1, BcBbBb = 1, BbBbBb = 1, BbBbDb = 1…

在每个时间序列获得直方图 H 之后,建立 TrSAX 字矩阵 M,其中每一行对应于时间序列数据,每一列对应于 TrSAX 字。表 2 给出了一个时间序列的 BOTS 表示形式的虚拟示例。

对于获得的矩阵 M,时间序列相似性度量为两行中列的累加平方值的平方根。给定矩阵 $M_{m\times n}$, $R_i = [m_{i1}, m_{i2}, \cdots, m_{in}]$ 和 $R_j = [m_{j1}, m_{i2}, \cdots, m_{jn}]$ 为矩阵的第 i 行和第 j 行。

表 2　时间序列的 BOTS 表示形式的虚拟示例

Table 2　Visual example of BOTS representation for time series

时间序列号	AaAaAa	AaAaAb	…	EaEaEa	…	EdEdEd
1	2	2	…	13	…	0
2	3	1	…	0	…	0
3	1	2	…	12	…	0
⋮	⋮	⋮		⋮	⋮	⋮
m	2	2	…	14	…	0

两者从时间序列 T_i 和 T_j 转换而来,由 BOTS 表示,BOTS 距离定义为

$$D(T_i, T_j) = \mathrm{dist}(R_i, R_j) \tag{6}$$

式中:

$$\mathrm{dist}(R_i, R_j) = \sqrt{(m_{i1} - m_{j1})^2 + (m_{i2} - m_{j2})^2 + \cdots + (m_{in} - m_{jn})^2} \tag{7}$$

5　实验分析

5.1　实验数据集

目前的卫星遥测数据各周期内的正常模式、异常模式、故障模式均没有明确资料可参考,因而选用具有卫星遥测参数数据类似特性的公开数据集进行验证。从 UCR 公开数据集中筛选出与卫星遥测数据中的角度序列类似的 SonyAIBORobotSurface 数据集,与卫星遥测数据中的转速序列类似的 Fish 数据集,与卫星遥测数据中的电流序列类似的 FaceFour 数据集进行实验验证,3 个时间序列数据集的信息如表 3 所示。

表 3　时间序列数据集详细信息

Table 3　Details on time series datasets

数据集编号	名称	训练集/测试集	类别数	序列长度
1	SonyAIBO RobotSurface	20/601	2	70
2	Fish	175/175	7	115
3	FaceFour	24/88	4	116

5.2　参数训练

每个数据集包含一个训练集和一个测试集。采用训练集训练参数。如 4.1 节所述,BOTS 分类器需要 4 个参数。

首先,使用网格搜索方法结合交叉验证从训练集中训练了这些参数。使用固定的大写字母数 $\gamma = 5$ 和字母符号数 $\varphi = 4$。对于段的数量 m,本文选择{3,4,5}。对于截长度 ω,都从 16 开始,每次加倍,直到小于或等于 $n/2$。训练算法如下所示。

训练阶段的目的是找到一组参数值,使得训练的分类准确率最大。如果 2 组参数具有相同的分类准确率,则选择较小的参数。在训练阶段的最后,获得了训练数据集的最佳参数值集合{γ, φ, m, ω},可直接用在测试集上进行分类。

算法 1　BOTS 分类算法。

```
From sk. learn. model_selection import cross_val_score
colLengh = len( trainInput [0])
end = int( math. log( colLengh/2,2))
windowList = list( map( int, np. logspace( 4, end, end - 3, base = 2)))   #截长度从 16 开始,每次加倍,直到小于或等于 n/2
wordList = [3,4,5]   #段的数量
tmpErrorRate, optimalWindow, optimalWord =
```

```
1,None,None
        for window in windowList:
            for word in wordList:
                testErrorRates =
    cross_val_score(trainAndPredict(window,
word,trainInput,trainLabels,testInput,testLabels),
X_trainval,Y_trainval,cv=5)    #5 折交叉验证
                testErrorRate = testErrorRates.
mean()    #取平均数
                logRecord.append([window,
word,testErrorRate])
                tmpErrorRate,optimalWindow,op-
timalWord = testErrorRate,window,word    #得到最
优的截长度、段的数量
```

5.3 TrSAX 各项参数的影响

所有的参数值集合在测试集上运行实验，分类错误率结果及运行时间如图8和图9所示。

如图8所示，段的数量 m 从3增加到4，分类错误率随着截长的变化，出现非线性的变化。但是当 m 值为5时，分类错误率明显增加。m 的值对分类错误率有影响，但没有一定规律。通常，对于本文中的大多数数据集的实验，值为3或4效果很好。

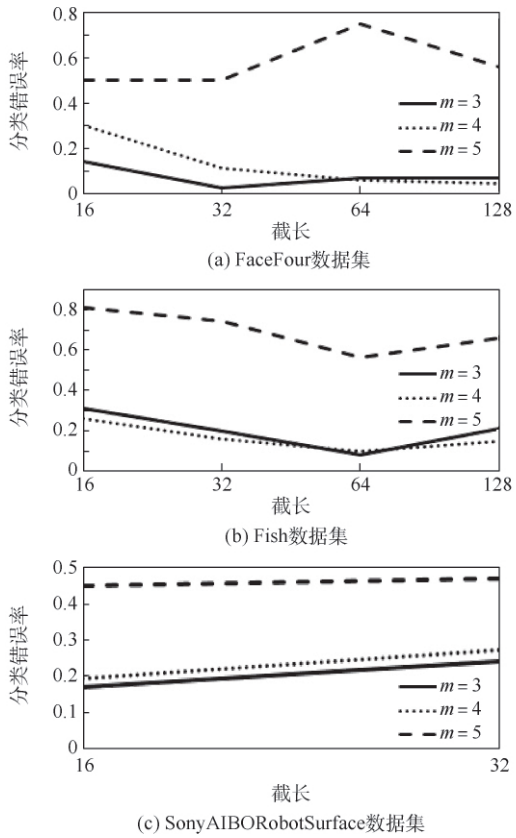

图8 三个数据集的分类错误率结果

Fig. 8 Classification Error rate results for three datasets

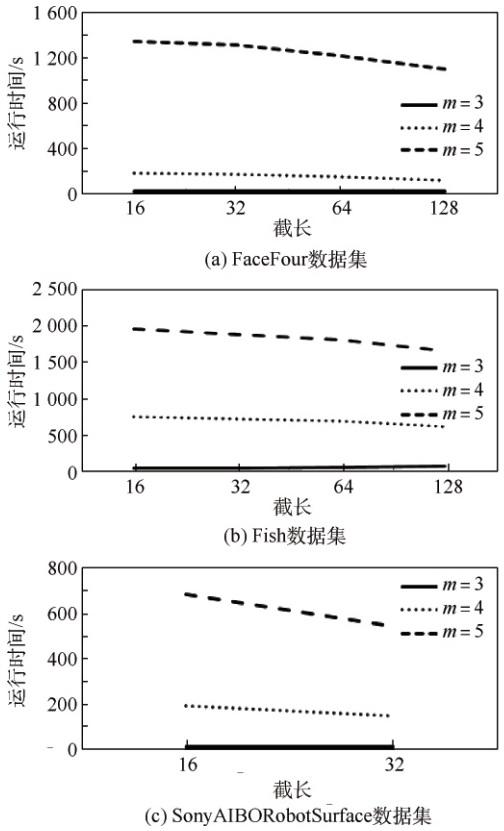

图9 三个数据集的运行时间结果

Fig. 9 Running time results for three datasets

如图9所示，段的数量 m 对运行时间的影响有2个规则。首先，m 从3增加到5，运行时间明显增加。原因是TrSAX词汇量为 $s=20^m$，随 m 的增加呈指数增长，并且与运行时间成正相关。另外，在 m 相同的情况下，运行时间随截长的增加而减少，并且当字典大小较大时，这种情况更加明显。这是因为词汇量大且稀疏，并且截长 ω 的增加减少了映射单词的数量，同时减小了所获得的矩阵 M 的大小，从而减少了运行时间。

5.4 对比算法及实验结果

由于本文方法旨在通过增加趋势信息来提高SAX的分类准确率。因此，将与1-NN SAX(或BOP)进行比较。另外，选择了ESAX和TSAX作为比较算法。如引言所述，基于ED和DTW建立的1-NN分类器在短时间序列上具有很高的代表性和竞争力。因此，将本文提出的BOTS与这5种算法进行比较，如表4所示。

通过对3个数据集进行预定义分类的实验，验证了BOTS的有效性。实验结果显示，BOTS分类算法在3个数据集中分类错误率都是最低，表现最佳。在平均排名方面，基本顺序是BOTS > 1-NN TSAX > 1-NN SAX > 1-NN DTW > 1-NN ED > 1-NN ESAX。

表 4　不同表示算法的分类结果

Table 4　Classification results of different representation algorithms

数据集	1-NN ED	1-NN DTW	1-NN SAX	1-NN ESAX	1-NN TSAX	BOTS
SongAIBO Robot	0.215	0.199	0.236	0.217	0.187	0.171
Fish	0.198	0.237	0.109	0.469	0.192	0.080
FaceFour	0.222	0.151	0.053	0.182	0.06	0.023
平均秩	4.67	4	3.33	5.33	2.67	1

在所选择的 3 个数据集中,1-NN DTW 和 1-NN ED 表现较好,略低于 1-NN SAX,优于 1-NN ESAX。而 BOTS 和 1-NN TSAX 表现优于 1-NN SAX,而 1-NN ESAX 表现排名最差,为第 6。说明趋势信息对降低分类错误率影响明显,对 SAX 增加合适的趋势信息可以明显降低分类错误率,而不合理的趋势信息则会明显增加分类错误率。

与 SAX 相比,TrSAX 的优点在于: 它可以通过每个段的趋势特征描述具有相同均值但趋势不同或趋势方向相同,但斜率有差异的某些模式(如缓增或速增)来提高分类的准确性。另外,对于某种时间序列,趋势信息对于专家而言非常重要,并且许多决策是通过趋势分析确定的。

TrSAX 也有一些缺点: 它仅是通过 0°和 ±45° 3 个角度间隔值简单地将趋势特征区分为速增、缓增、水平、缓降和速降 5 个状态,这对于一些数据集有用。对于一些特殊的数据集,不同的角度间隔会使得分类准确率不同。如何合理确定角度间隔值的问题仍未解决。

6　结　论

1) 本文提出了一种基于趋势符号聚合近似的时间序列表示方法 TrSAX,可以将原始时间序列转换为矩阵结构,该矩阵结构可以使用最小二乘法结合 SAX 表示来描述每个时间序列段的平均值和斜率值。

2) 对文献[21]中的"风云三号"卫星遥测数据进行了分析,得出角度序列、转速序列、电流序列 3 种遥测数据具有明显的周期性,以角度测试时间序列的幅角突变点为标识进行分段,进行叠加绘图,可以得到耦合度较高的周期曲线。在 UCR 公共数据集中筛选出与角度序列、转速序列、电流序列类似的 3 个数据集进行了实验验证。

3) 通过实验分析段的数量 m,截长度 ω 的各种值对 BOTS 分类性能的影响。期望较大 m 和较小的 ω 来描述原始时间序列的尽可能多的细节,并为时间序列分类带来更好的结果。但是,实验结果证明这一期望没有实现。随着 m 的增加或 ω 的减少,分类错误率不会持续下降。为了确定 m 和 ω 的最佳组合,需要在提供的训练数据集上进行训练。此外,m 对分类错误率更敏感,并且 m 比 ω 具有更大的影响。对于大多数数据集,m 等于 3 或 4 可以产生令人满意的结果。

4) 对比实验结果显示,与卫星遥测参数中角度序列、转速序列、电流序列类似的 3 个公开数据集中,BOTS 分类错误率明显低于其他 5 种分类方法,对未来卫星模拟量遥测参数的分类提供了一种新的思路。

在未来的研究中,将研究参数 m 和 ω 的设置,以加速和简化参数的确定,同时需要对角度间隔值进行进一步的优化。此外,还将探索基于 TrSAX 的其他任务,如聚类以及相关性分析等。

参考文献(References)

[1] 杨海民,潘志松,白玮. 时间序列预测方法综述[J]. 计算机科学,2019,46(1):21-28.
YANG H M,PAN Z S,BAI W. Review of time series prediction methods[J]. Computer Science,2019,46(1):21-28(in Chinese).

[2] 史欣田,庞景月,张新,等. 基于集成极限学习机的卫星大数据分析[J]. 仪器仪表学报,2018,39(12):81-91.
SHI X T,PANG J Y,ZHANG X,et al. Satellite big data analysis based on bagging extreme learning machine[J]. Chinese Journal of Scientific Instrument,2018,39(12):81-91(in Chinese).

[3] 彭喜元,庞景月,彭宇,等. 航天器遥测数据异常检测综述[J]. 仪器仪表学报,2016,37(9):1929-1945.
PENG X Y,PANG J Y,PENG Y,et al. Review on anomaly detection of spacecraft telemetry data[J]. Chinese Journal of Scientific Instrument,2016,37(9):1929-1945(in Chinese).

[4] YANG T,CHEN B,GAO Y,et al. Data mining-based fault detection and prediction methods for in-orbit satellite[C]//IEEE International Conference on Measurement,Information and Control. Piscataway: IEEE Press,2013:805-808.

[5] 肇刚,李言俊. 基于时间序列数据挖掘的航天器故障诊断方法[J]. 飞行器测控学报,2010,29(3):1-5.
ZHAO G,LI Y J. Spacecraft fault diagnosis method based on time series data mining[J]. Journal of Spacecraft TT & C Technology,2010,29(3):1-5(in Chinese).

[6] 鲍军鹏,杨科,周静. 卫星时序数据挖掘节点级并行与优化方法[J]. 北京航空航天大学学报,2018,44(12):2470-2478.
BAO J P,YANG K,ZHOU J. Node level parallel and optimiza-

tion method of satellite time serial data mining [J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(12): 2470-2478(in Chinese).

[7] 张弓,翟君武,杨海峰. 导航卫星遥测数据趋势预测技术研究[J]. 航天器工程,2017,3(3):74-81.
ZHANG G, ZHAI J W, YANG H F. Research on telemetry data tendency prognosis for navigation satellite [J]. Spacecraft Engineering, 2017, 3(3): 74-81(in Chinese).

[8] WAN Y, SI Y W. A hidden semi-Markov model for chart pattern matching in financial time series [J]. Soft Computing, 2017, 22(3): 1-20.

[9] MUEEN A, KEOGH E, YOUNG N E. Logical-Shapelets: An expressive primitive for time series classification [C]// ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2011: 1154-1162.

[10] GAO Z K, CAI Q, YANG Y X, et al. Multiscale limited penetrable horizontal visibility graph for analyzing nonlinear time series [J]. Scientific Reports, 2016, 6(1): 35622.

[11] XI X, KEOGH E, SHELTON C, et al. Fast time series classification using numerosity reduction [C]// International Conference On Machine Learning, 2006: 1033-1040.

[12] 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.

[13] RAKESH A, CHRISTOS F, ARUN S. Efficient similarity search in sequence databases [C]// Foundations of Data Organization and Algorithms. Berlin: Springer, 1993: 69-84.

[14] CHEN Q, CHEN L, LIAN X, et al. Indexable PLA for efficient similarity search [C]// VLDB Endowment in Proceedings of the 33rd International Conference on Very Large Data Bases. 2007: 435-446.

[15] LIN J, KEOGH E, LI W, et al. Experiencing SAX: A novel symbolic representation of time series [J]. Data Mining & Knowledge Discovery, 2007, 15(2): 107-144.

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

[17] LIN J, KHADE R, LI Y. Rotation-invariant similarity in time series using bag-of-patterns representation [J]. Journal of Intelligent Information Systems, 2012, 39(2): 287-315.

[18] PHAM N D, LE Q L, DANG T K. Two novel adaptive symbolic representations for similarity search in time series databases [C]// Proceedings of the 12th Asia-Pacific Web Conference (APWeb). Piscataway: IEEE Press, 2010: 181-187.

[19] LKHAGVA B, SUZUKI Y, KAWAGOE K. New time series data representation ESAX for financial applications [C]// International Conference on Data Engineering Workshops. Piscataway: IEEE Press, 2006: 17-22.

[20] ZHANG K, LI Y, CHAI Y, et al. Trend-based symbolic aggregate approximation for time series representation [C]// 2018 Chinese Control and Decision Conference (CCDC). Piscataway: IEEE Press, 2018: 2234-2240.

[21] 陈静. 卫星遥测数据的时间序列相似性度量方法研究[D]. 哈尔滨:哈尔滨工业大学,2015:22-23.
CHEN J. Similarity measure of time series for satellite telemetry data [D]. Harbin: Harbin Institute of Technology, 2015: 22-23(in Chinese).

作者简介:

阮辉 男,博士研究生。主要研究方向:卫星故障诊断、数据挖掘、数字信号处理。

刘雷 男,硕士,高级工程师。主要研究方向:飞行器电气系统。

胡晓光 女,博士,教授,博士生导师。主要研究方向:图像处理、故障诊断、嵌入式测试系统和智能电网。

Satellite time series data classification method based on trend symbolic aggregation approximation

RUAN Hui[1], LIU Lei[2], HU Xiaoguang[1*]

(1. School of Automation Science and Electrical Engineering, Beihang University, Beijing 100083, China;
2. Beijing Electro-Mechanical Engineering Institute, Beijing 100074, China)

Abstract: As the main symbolic representation method widely used in time series data mining, the Symbolic Aggregation Approximation (SAX) uses the mean value of segments as the symbolic representation. Since it is impossible to distinguish different time series that have different trends but the same mean value, it may lead to incorrect classification. This paper presents an improved symbol representation—Trend Symbol Aggregation Approximation (TrSAX), which integrates SAX and least squares method to describe the mean and slope value of the time series, and constructs the BOTS classifier. In addition, this paper analyzes the angle sequence, rotation speed sequence, and current sequence in the satellite analog telemetry time series data, and selects three datasets similar to these three sequences from the UCR public dataset for classification experiment verification. They are compared with the 1-NN classification methods using SAX, two improved SAX, classic Euclidean Distance (ED) and Dynamic Time Warping (DTW). The results show that the classification error rate of the proposed BOTS classification method is significantly lower than the other five classification methods.

Keywords: satellite telemetry data; time series; symbolic representation; time series classification; anomaly detection

Received: 2020-07-12; **Accepted:** 2020-08-07; **Published online:** 2020-08-18 10:02

URL: kns.cnki.net/kcms/detail/11.2625.V.20200817.1820.002.html

Foundation items: National Natural Science Foundation of China (51807003); National Defense Basic Scientific Research Program of China (JKCY2016204A102)

* **Corresponding author**. E-mail: xiaoguang@buaa.edu.cn