基于 PAA 的时间序列早期分类

马超红 翁小清
(河北经贸大学信息技术学院 石家庄 050061)

摘 要 在时间序列数据挖掘领域,时间序列的早期分类越来越受到人们的重视,由于时间序列的长度(也称为维数)较大,在早期分类的实际应用中选择合适的维数约简方法非常重要,因此提出一种基于分段聚合近似(PAA)的时间序列早期分类方法。首先运用 PAA 对时间序列样本进行维数约简,然后在低维空间对样本进行早期分类,在 43 个时间序列数据集上的实验结果表明,所提方法在准确率、早期性、可靠性等方面优于已有方法。

关键词 时间序列,早期分类,维数约简,分段聚合近似
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.050

Early Classification of Time Series Based on Piecewise Aggregate Approximation

MA Chao-hong WENG Xiao-qing
(School of Information Technology, Hebei University of Economics and Business, Shijiazhuang 050061, China)

Abstract Early classification on time series is more and more significant in the field of time series data ming. As the high dimension of time series data, it is of highly necessary to choose an efficient and appreciate dimensionality reduction method in the practical application of early classification on time series. Thus, this paper aimed at applying piecewise aggregate approximation to time series data, and then implemented early classification in lower dimension. In addition, through making comparison with some existing methods, the experiments were carried on in forty-three datasets. The experimental result indicates that this proposal is better than other existing methods in accuracy, earliness and reliability.

Keywords Time series, Early classification, Dimensionality reduction, Piecewise aggregate approximation

1 引言

时间序列[1]指按时间顺序有次序排列的一组数据,任何实值型有次序的序列都可以作为时间序列来处理。时间序列分类是将待测样本分配到预先定义好的类别中,在医学诊断、灾害预测、入侵检测、过程控制、道路交通等领域得到了广泛的应用和研究。时间序列早期分类[1]是在满足预测质量(准确率)的前提下尽可能早地进行分类,即不使用全部长度的序列样本,只根据序列的前端部分做出分类。时间序列早期分类在时效性要求高的应用领域中具有重要意义。对于时间序列,由于样本随着时间不断增长,因此约简时间序列的长度即序列样本中的元素个数(文献[3]称之为维数)是实际分类中的一个巨大难题。Xing 等[5]提出了一种基于 1 近邻的早期分类方法(Early Classification on Time Series, ECTS),ECTS 直接对原始数据进行早期分类,没有对数据进行维数约减。在目前的早期分类算法中,只有 Parrish 等[2]提出时间序列早期分类方法(RelClass),其采用局部可鉴别的高斯方法(Local Discriminative Gaussian, LDG)从时间序列训练集中找到一个转换矩阵,将时间序列映射到低维空间,但在早期分类实际应用中,如何对测试样本的部分数据进行维数约简,并没有得到解决。本文在 ECTS 方法的基础上,提出了一种基于分段聚合近似[3](Piecewise Aggregate Approximation, PAA)的早期分类方法(PAA_ECTS),先使用 PAA 对原始数据进行维数约简,再在低维空间进行早期分类。实验结果表明,在准确率、早期性、可靠性等方面,PAA_ECTS 方法都优于已有的早期分类算法。

本文第 2 节介绍研究背景及相关工作;第 3 节介绍提出的时间序列早期分类方法 PAA_ECTS;第 4 节通过实验将 PAA_ECTS 与现有早期分类方法进行比较,并采用威尔克森符号秩检验(Wilcoxon Signed Ranks Test)对实验结果进行评估;最后总结全文并展望下一步工作。

2 基本概念与相关工作

2.1 基本概念

本节给出了本文涉及的相关定义和符号解释,并对分段聚合近似做了相关介绍。

定义 1(单变量时间序列) $X=\{x_1, x_2, \cdots, x_n\}$ 表示长度为 n 的一条时间序列,其中 x_i $(1\leqslant i\leqslant n)$ 表示时刻 i 对应的数值。

到稿日期:2016-12-02 返修日期:2017-02-27

马超红(1994—),女,硕士生,CCF 会员,主要研究方向为数据挖掘与信息检索,E-mail:chaohma@126.com;**翁小清**(1965—),男,博士,教授,主要研究方向为数据挖掘,E-mail:xqweng@126.com(通信作者)。

定义 2(反向最近邻集合) 时间序列 t 的前缀 $t(1,l)$ 在前缀空间 R^l 中的反向最近邻集合 $RNN^l(t)$[5] 为:

$$RNN^l(t)=\{t'\in T \mid t\in NN^l(t')\}$$

其中,t 和 t' 是训练集 T 中的样本,$NN^l(t')$ 是 t' 在 R^l 中的最近邻。

定义 3(最小预测长度) 最小预测长度(Minimum Prediction Length,MPL)[5] 表示在序列长度为 n 的训练集 T 中,对于 $t\in T$,$MPL(t)=k$,当且仅当对于任意的 $l(k\leqslant l\leqslant n)$,存在:1)$RNN^l(t)=RNN^n(t)\neq\emptyset$;2)$RNN^{k-1}(t)\neq RNN^n(t)$。

特别地,当 $RNN^n(t)=\emptyset$ 时,$MPL(t)=n$。

定义 4(簇 S 的最小预测长度) 簇 S 的最小预测长度(MPLs of cluster,MPL(S))[5] 中,S 为一个簇,$MPL(S)=k$,当且仅当对于所有的 $l>k$ 时,满足下列条件:1)$RNN^l(S)=RNN^L(S)$;2)在前缀空间 R^l 中,S 是 1 近邻连续的;3)对于任意的 $l=k-1$,条件 1)和条件 2)不能同时满足。$RNN^l(S)$ 的计算公式为:$RNN^l(S)=\bigcup_{s\in S}RNN^l(S)\backslash S$。

定义 5(准确性) 准确性(Accuracy)[5,7] 即将对测试集进行分类时分类结果正确的概率,其按式(1)进行计算,其中 $\hat{C}_i$ 表示测试样本的实际类标号,C_i 表示样本在进行早期分类即序列长度为 t_i 时分类器输出的类标号,N 代表训练集样本个数。对于 $I(\cdot)$,当括号内条件满足时其值为 1,否则为 0。

$$Accuracy=\frac{1}{N}\sum_{i=1}^{N}I(\hat{C}_i=C_i) \tag{1}$$

定义 6(早期性) 用测试样本得出分类结果时所需的时间序列长度来衡量早期性(Earliness)[7],本文取所需样本长度 t_i 占全长序列长度 L 的百分比,如式(2)所示:

$$Earliness=\frac{1}{N}\sum_{i=1}^{N}\frac{t_i}{L}\times 100 \tag{2}$$

定义 7(可靠性) 可靠性(Reliability)[7] 指采用早期分类方法即较短的序列长度 t_i 的分类结果与采用全长序列长度 L 的分类结果的比较,计算公式如式(3)所示,C_{i_full} 表示分类器输入为全长序列时输出的类标号。

$$Reliability=\frac{1}{N}\sum_{i=1}^{N}I(C_{i\_full}=C_i) \tag{3}$$

分段聚合近似(PAA)[3] 是一种基于分段序列平均值的特征提取算法。PAA 先将时间序列划分为等长的分段,然后计算这些分段的平均值,用平均值表示该分段中所有数据的值。设时间序列 $X=\{x_1,x_2,\cdots,x_n\}$ 的长度为 n,使用 PAA 算法后长度降至 d 维,可用 $\overline{X}=\{\overline{x_1},\cdots,\overline{x_i},\cdots,\overline{x_d}\}$ 表示 X,其中:

$$\overline{x_i}=\frac{d}{n}\sum_{j=\frac{n}{d}(i-1)+1}^{\frac{n}{d}i}x_j \tag{4}$$

进行 PAA 降维[4] 后,在较低的维数空间中进行后续计算,降低了时间复杂度;另外,对于每一分段,用其平均值替代,在不同程度上对噪声进行了平滑,从而可以提高分类器的性能。但 PAA 也存在弊端,由于每个分段用其平均值来表示,因此会丢失时间序列中一些比较关键的信息,如极大值、极小值和其他一些重要形态等。另外,PAA 降维后序列的信息取决于所降的维数 d,d 越小,PAA 所表示的时间序列越粗糙,丢失的信息越多,维数约简的程度越大;反之,PAA 所表示的时间序列越精细,丢失的信息越少,维数约简的程度越小。因此在使用 PAA 进行降维时,要考虑对降维后样本的质量与分段个数 d 的权衡。

PAA 的时间复杂度是线性的,实现简单,适合于对任意长度的样本进行约简。根据分段数,计算每段的平均值,其也适合于早期分类的实际应用。实验表明,其在保持准确率较高的前提下,能尽可能早地对序列进行分类,实现了可靠的时间序列早期分类。

2.2 相关工作

时间序列的早期分类方法[1] 大致分为 3 类:基于原始数据的分类方法、基于特征的分类方法和基于模型的分类方法。Xing 等[5-6] 于 2009 年将 1NN 分类器用于时间序列的早期分类,提出了最小预测长度 MPL 的概念和时间序列早期分类方法 ECTS,ECTS 适用于 1NN 分类器有效的情况,既能保证分类准确率,又能实现早期分类。ECTS 采用单链接 MLHC(Multi-level Hierarchical Clustering)在不同的前缀空间进行聚类,并且计算每一个聚类后的簇的最小预测长度 MPL。Xing 等还提出了 Relaxed ECTS,通过 relaxed MPL 避免在原 ECTS 方法下由于处于决策边界而没有被分类的时间序列,从而提高了分类器的整体稳定性。Mori 等[7] 提出了基于概率分类器的 ECDIRE,该方法在训练阶段分析类之间的区别性,并且选择一系列的时刻进行分类,所得结果的准确率超过了预先给定的阈值,从而能避免过早做出预测。Parrish 等[2] 提出采用不完整数据来进行可靠分类,设置了 τ 值,当概率大于或等于 τ 时,利用不完整数据作出预测,确保采用不完整数据进行分类的结果与利用完整数据进行分类的结果尽可能相似,该方法在线性或二次判别式(LDA&QDA)分类器中采用最优和实用的决策规则来有效并正确地分类不完整数据。Xing 等[8] 于 2011 年在早期分类中提取具有可解释性的特征(Local Shapelets),针对单变量时间序列,提出了 best match distance(BMD)和 BMD-lists 的概念,并分别使用核密度估计方法(Kernel Density Estimation)和切比雪夫不等式(Chebyshev's Ineqaulity)来学习 local shapelet $f=(s,\delta,c)$ 中的距离阈值 δ。通过计算每一个 shapelet 的效用值 Utility 排序,根据 Utility 值来选择用于分类的 shapelets。

在已有的早期分类算法中,只有 Parrish 等[2] 提出的 RelClass 使用 LDG 对原始数据进行维数约简,然而,RelClass 没有对测试样本的部分数据如何进行维数约简给出解决办法,因此其无法用于早期分类的实际应用。而本文提出的基于 PAA 的早期分类方法对测试样本的部分数据分段计算平均值,以达到维数约简的目的,在早期分类的实际应用中非常简便。

3 基于 PAA 的时间序列早期分类算法

本文提出基于 PAA 的时间序列早期分类算法(PAA_

ECTS)。首先使用 PAA 对时间序列进行维数约简，然后使用 Xing 等[5-6]提出的 ECTS 方法对维数约简后的数据进行早期分类。在算法中，第一步，根据 PAA 维数确定分段长度，将时间序列划分为等长分段，并将该序列用每段的平均值组成的向量来近似表示，得到维数约简后的数据；第二步，即测试阶段，计算维数约简后的数据集的每一个样本的 $MPL(t)$；第三步，即早期分类，根据第一步确定的分段长度，对待分类样本分段计算其平均值，不足一个分段长度的继续等待，直到数据够一个分段长度，然后计算其近邻，当满足分类条件时做出决策，分类终止。

算法 1　PAA_ECTS 早期分类算法

输入：训练集 TRAIN，测试集 TEST，PAA 维数 d

输出：Accuracy，Earliness，Reliability

Step1　将 TRAIN 中的每条序列 X_i 等分为 d 段，并计算分段长度及每段数据的平均值；

Step2　以 d 段平均值组成的向量作为 X_i 的表示，得到在 d 维空间的集合 PTRAIN；

Step3　对于降维后的数据集 PTRAIN，在所有的前缀空间中求得每个样本的 1 近邻；

Step4　采用单链接 MLHC[9]对 PTRAIN 进行聚类，得到若干子簇 S；

Step5　计算每个叶子簇的 MPL，并将该值赋给当前簇中每个样本 t 的 MPL(t)；

Step6　迭代计算所有簇中最近的簇对(S_1，S_2)，每次迭代将 S_1 和 S_2 合并为一个父簇；

Step7　对于属于簇 S 中的样本 t，当 MPL(S)＜MPL(t)时，将 MPL(S)的值赋给 MPL(t)；

Step8　当所有的样本 t 的 MPL(t)均小于或等于包含该样本簇的 MPL(S)时，迭代终止；

Step9　将 TEST 中的每个时间序列 s 等分为在 Step1 中所得的分段长度，计算每段数据的平均值，以每段平均值组成的向量作为 s 的表示；

Step10　在时刻 i，如果其近邻 $NN^i(s)$的 MPL 为 i，则返回 $NN^i(s)$的类别号作为 s 的类标号；

Step11　如果在时刻 i 不存在这样的 $NN^i(s)$，则分类器认为在当前时刻不能做出可靠决策，继续等待待分类序列的下一段到达分类器；

Step12　计算 TEST 集的 Accuracy，Earliness，Reliability。

PAA_ECTS 算法分 3 个阶段，Step1 和 Step2 为降维阶段，PAA 算法是线性的，其复杂度为 $O(n)$，n 为序列长度；Step3－Step8 是训练阶段，需要在不同的前缀空间中采用单链接 MLHC 聚类得到若干子簇，计算样本间的相似性，确定样本的最近邻和反向最近邻，时间复杂度[5]为 $O(|T|^3 \cdot n)$，其中 $|T|$ 为训练集样本个数；Step9－Step12 是分类阶段，计算训练集样本在测试集样本内的 1 近邻，计算时间复杂度为 $O(|D| \cdot n)$，其中 $|D|$ 为测试集样本个数。综上，PAA_ECTS 的总体复杂度为 $O(|T|^3 \cdot n+|D| \cdot n)$。

4　实验

本节从准确性、早期性和可靠性三方面评估 PAA_ECTS 的性能。所用数据集为 UCR 档案库[10]，UCR 储存了大量公开可用的合成及真实的时间序列数据库。所比较的方法包括 ECTS[5-6]，EDSC[8]，RelClass[2] 和 ECDIRE[7]。其中，EDSC 的阈值设为 2.5，RelClass 选用 Gaussian Naive Bayes box 方法，设 $\tau=0.5$，设置 ECTS 的支持度 $support=0$。

4.1　数据集描述

在 43 个时间序列数据集上[10]进行测试实验，表 1 列出了这 43 个时间序列数据集的主要特征，包括数据集名称、类别个数、训练集样本总数、测试集样本总数和样本的长度。这些时间序列数据集来自医学、工业、图像识别、视频、生物等领域。

表 1　数据集描述

Table 1　Description of datasets

编号	数据集名称	类别个数	训练集样本总数	测试集样本总数	时间序列长度
1	Synthetic_Control	6	300	300	60
2	Gun_Point	2	50	150	150
3	CBF	3	30	900	128
4	Face_All	14	560	169	131
5	OSU Leaf	6	200	242	427
6	Swedish_Leaf	15	500	625	128
7	50Words	50	450	455	270
8	Trace	4	100	100	275
9	Two_Patterns	4	100	400	128
10	Wafer	2	100	617	152
11	Face(four)	4	24	88	350
12	Lightning_2	2	60	61	637
13	Lightning_7	7	70	73	319
14	ECG	2	100	100	96
15	Adiac	37	390	391	176
16	Yoga	2	300	300	426
17	Fish	7	175	175	463
18	WordSynonyms	25	267	638	270
19	Beef	5	30	30	470
20	Coffee	2	28	28	286
21	OliveOil	4	30	30	570
22	CinC_ECG_torso	4	40	138	163
23	ChlorineConcentration	3	467	384	166
24	DiatomSizeReduction	4	16	306	345
25	ECGFiveDays	2	23	861	136
26	FacesUCR	14	200	205	131
27	Haptics	5	155	308	109
28	InlineSkate	7	100	550	188
29	ItalyPowerDemand	2	67	102	24
30	MALLAT	8	55	234	102
31	MediaImages	10	381	760	99
32	MoteStrain	2	20	125	84
33	SonyAIBORobot	2	27	953	65
34	SonyAIBORobot Surface	2	20	601	70
35	StarLightCurves	3	100	823	102
36	Symbols	6	25	995	398
37	TwoLeadECG	2	23	113	82
38	Cricket_X	12	390	390	300
39	Cricket_Y	12	390	390	300
40	Cricket_Z	12	390	390	300
41	uWaveGestureLibrary_X	8	890	358	315
42	uWaveGestureLibrary_Y	8	890	358	315
43	uWaveGestureLibrary_Z	8	890	358	315

4.2　性能比较

各种方法的 Accuracy 结果如表 2 所列，其中 Full_1NN 表示 ECTS 分类器使用完整数据集分类时对应的 Accuracy；各种方法的 Earliness 结果如表 3 所列。表 2 中的 PAA_

ECTS 的实验结果中 Accuracy 以及相应的 PAA 维数 d(括号中的内容表示该结果对应的 PAA 维数)是最高的,该参数的选取方法如下:使用 PAA_ECTS 在某个数据集上进行多次实验,PAA 维数 d 的取值范围为从 2 到该数据集时间序列样本的长度 n,选取分类性能最好的维数 d;表 3、表 4 中 PAA_ECTS 的结果也是在表 2 对应的 PAA 维数的基础上计算所得。需要指出的是,ECDIRE,RelClass 和 EDSC 的 Accuracy 与 Earliness 结果均取自文献[7]。

为了对早期分类的实验结果进行显著差异的判断,本文采用 Wilcoxon 符号秩检验,该方法是非参数统计方法,对数据的分布没有要求,概率 p 值小于 0.05 说明差异显著,差异显著性见表 5。

表 2 PAA_ECTS 与 ECDIRE,RelClass,ECTS,EDSC,Full_1NN 的 Accuracy 结果比较

Table 2 Comparison of Accuracy results of PAA_ECTS, ECDIRE, RelClass, ECTS, EDSC and Full_1NN

数据集	ECDIRE	RelClass	ECTS	EDSC	Full_1NN	PAA_ECTS
50words	53	66	58.24	48	63.08	64.40(13)
Adiac	55	63	60.61	16	61.13	61.64(148)
Beef	50	57	53.33	23	53.33	53.33(68)
CBF	89	64	85.22	84	85.22	96.67(10)
Chlorine_concentration	56	82	61.90	52	65.00	62.40(139)
CinC_ECG_torso	81	85	87.46	55	89.71	90.00(75)
Coffee	96	89	78.57	75.00	75.00	82.14(230)
Cricket_X	57	61	56.92	52	57.44	65.64(26)
Cricket_Y	63	68	63.33	57	64.36	67.44(14)
Cricket_Z	60	66	58.97	0	62.05	66.67(14)
Diatom_size_reduction	80	94	80.07	85	93.46	89.54(3)
ECG200	91	89.00	89.00	85	88.00	91.00(55)
ECG_five_days	60	52	62.49	74	79.67	83.04(11)
Face_All	87	69	73.61	66	71.36	76.45(92)
FaceFour	61	83	77.27	75	78.41	88.64(227)
FacesUCR	74	77	71.8	63	76.93	76.34(33)
Fish	81	79	74.86	68	78.29	79.43(259)
Gun_Point	87	91	86.67	94	91.33	92.00(6)
Haptics	44	41	37.34	34	37.01	39.61(282)
Inline_skate	26	27	32.73	18	34.18	33.64(26)
Italy_power_semand	93	85	93.97	82	95.53	93.97(24)
Lighting2	54	62	70.49	80	75.41	86.89(217)
Lighting7	48	68	57.53	67	57.53	73.97(10)
MALLAT	78	73	84.61	59	91.43	93.43(20)
Medical_images	74	67	67.76	60	68.42	68.82(48)
Mote_strain	80	58	87.86	78	87.86	87.86(84)
Olive_oil	40	77	90.00	60	86.67	90.00(55)
OSU_leaf	52	48	48.76	56	51.65	54.13(248)
Sony_AIBO_robot_surface	83	79	68.72	80	69.55	80.37(15)
Sony_AIBO_robot_surfaceII	74	88	84.47	81	85.94	87.93(17)
Star_light_curves	95	95	85.21	—	84.88	89.29(9)
Swedh_leaf	87	83	78.88	47	78.88	82.56(21)
Symbols	81	71	82.61	51	89.95	85.23(22)
Synthetic_control	96	98	89.00	89	88.00	98.00(10)
Trace	77	86	74.00	80	76.00	76.00(198)
TwoLeadECG	81	72	73.49	88	74.71	75.24(48)
Two_Patterns	87	93	86.48	80	90.68	91.83(18)
uWaveGestureLibrary_X	77	75	73.12	54	73.93	73.67(12)
uWaveGestureLibrary_Y	70	68	63.32	37	66.16	65.83(14)
uWaveGestureLibrary_Z	71	71	64.68	52	64.96	65.89(14)
Wafer	97	99	99.24	99	99.55	99.29(109)
Words_synonyms	52	65	58.78	47	61.76	61.91(18)
Yoga	85	83	81.40	71	83.03	81.73(14)
平均值	71.7	73.65	72.44	62.43	74.59	77.3

从表 2 以及表 5 可以看到,PAA_ECTS 与 ECTS,ECDIRE 及 EDSC 的准确率之间的 Wilcoxon 符号秩检验的概率 p 值都小于 0.05,这说明 PAA_ECTS 的准确率显著优于 ECTS,ECDIRE 及 EDSC。在 43 个数据集上,PAA_ECTS 的平均准确率是 77.30%,而 ECTS,ECDIRE 及 EDSC 的平均准确率分别是 72.44%,71.70%,62.43%。原因在于采用 PAA 进行维数约简时计算的是每段的平均值,对时间序列中存在的噪声、离群点等进行了有效的平滑。PAA_ECTS 与 RelClass 的准确率之间的 Wilcoxon 符号秩检验的概率 p 值大于 0.05,说明 PAA_ECTS 的准确率虽然高于 RelClass 的准确率,但是差别不显著。在 43 个数据集上,PAA_ECTS 的平均准确率是 77.30%,RelClass 的平均准确率是 73.653%,然而 RelClass[2] 并没有解决对测试样本的部分数据进行维数约简的问题,所以它无法用于早期分类的实际应用,而文中提

出的 PAA_ECTS 方法可以非常容易地对测试样本的部分数据进行维数约简。

表 3　PAA_ECTS 与 ECDIRE，RelClass，ECTS 及 EDSC 的 Earliness 结果比较

Table 3　Comparison of Earliness results of PAA_ECTS，ECDIRE，RelClass，ECTS and EDSC

数据集	ECDIRE	RelClass	ECTS	EDSC	PAA_ECTS
50words	40.30	92.20	77.10	58.89	80.34
Adiac	38.54	96.04	64.05	84.55	54.62
Beef	67.78	25.7	77.67	93.61	72.13
CBF	28.55	23.08	71.67	31.85	64.39
Chlorine_concentration	14.42	97.59	67.01	33.33	59.17
CinC_ECG_torso	49.71	56.58	59.91	43.63	57.48
Coffee	82.14	38.44	84.93	54.23	75.35
Cricket_X	47.98	78.68	73.26	52.57	69.58
Cricket_Y	36.00	82.36	67.47	45.10	71.29
Cricket_Z	45.99	80.36	69.21	56.12	74.22
Diatom_size_reduction	24.26	33.49	14.88	27.04	76.91
ECG200	90.10	68.81	77.13	23.24	38.91
ECG_five_days	21.07	15.84	63.82	53.60	78.03
Face_All	56.49	96.27	68.25	38.94	48.24
FaceFour	22.31	34.22	89.51	47.98	46.33
FacesUCR	59.15	92.71	89.24	51.58	67.70
Fish	55.17	85.42	65.81	47.70	45.20
Gun_Point	32.37	71.33	46.92	45.58	66.67
Haptics	86.52	57.89	93.90	12.53	70.53
Inline_skate	33.83	87.31	86.42	46.69	87.67
Italy_power_semand	70.16	35.92	79.33	67.08	79.33
Lighting2	9.07	61.16	84.83	55.14	65.58
Lighting7	19.93	85.23	86.98	68.40	81.34
MALLAT	45.35	44.01	69.32	39.96	77.90
Medical_images	21.20	88.96	54.85	31.95	57.78
Mote_strain	12.10	90.94	84.86	38.08	84.86
Olive_oil	30.00	18.76	87.34	38.82	74.91
OSU_leaf	47.52	97.10	78.20	54.38	48.66
Sony_AIBO_robot_surface	62.26	57.70	68.49	47.03	61.60
Sony_AIBO_robot_surfaceII	17.66	70.86	55.81	35.51	44.48
Star_light_curves	53.10	90.02	82.83	—	85.67
Swedh_leaf	45.97	91.96	77.63	62.34	77.64
Symbols	45.33	45.82	46.23	60.25	64.53
Synthetic_control	61.92	71.54	89.96	50.81	88.53
Trace	41.75	77.82	51.98	38.63	46.00
TwoLeadECG	69.38	83.63	64.43	46.85	41.83
Two_Patterns	98.76	91.82	86.79	64.04	84.83
uWaveGestureLibrary_X	74.03	90.09	86.90	64.30	89.62
uWaveGestureLibrary_Y	97.09	81.96	86.91	70.14	87.82
uWaveGestureLibrary_Z	75.56	91.80	85.98	61.18	87.01
Wafer	10.87	30.75	44.38	27.99	28.08
Words_synonyms	66.23	91.40	83.40	65.66	88.58
Yoga	100	87.28	70.74	38.57	76.85
平均值	49.02	69.55	72.47	49.43	68.10

由表 3 以及表 5 的结果可以看出，PAA_ECTS 与 EDSC 和 ECDIRE 的早期性之间的 Wilcoxon 符号秩检验的概率 p 值都小于 0.05，这说明在早期性方面，EDSC 和 ECDIRE 明显优于 PAA_ECTS；在 43 个时间序列数据集上，EDSC 和 ECDIRE 的平均早期性分别为 49.43%和 49.02%，PAA_ECTS 的平均早期性为 68.10%；然而，它们都是在牺牲准确率的前提下实现的，PAA_ECTS 的平均准确率是 77.30%，而 EDSC 为 62.43%，ECDIRE 为 71.70%。PAA_ECTS 与 ECTS 的早期性之间的 Wilcoxon 符号秩检验的概率 p 值小于 0.05，这说明 PAA_ECTS 的早期性明显优于 ECTS；在 43 个时间序列数据集上，PAA_ECTS 的平均早期性为 68.10%，ECTS 为 72.47%。

PAA_ECTS 与 ECTS 的 Reliability 见表 4。对 ECTS 与 PAA_ECTS 的 Reliability 进行了 Wilcoxon 符号秩检验，统计量 z 值为 −3.2848，概率 p 值为 0.001（小于 0.05），PAA_ECTS 的可靠性显著优于 ECTS，在 43 个时间序列数据集上 PAA_ECTS 的平均可靠性为 92.61%，ECTS 为90.88%。

表 4　ECTS 与 PAA_ECTS 的 Reliability 实验结果比较

Table 4　Comparison of Reliability results of ECTS and PAA_ECTS

数据集	ECTS	PAA_ECTS
50words	87.25	89.89
Adiac	86.70	85.42
Beef	93.33	90.00
CBF	96.67	97.33
Chlorine_concentration	88.28	89.04
CinC_ECG_torso	93.41	95.87
Coffee	96.43	100.00
Cricket_X	87.95	87.44
Cricket_Y	83.08	84.62
Cricket_Z	83.59	87.69
Diatom_size	85.29	98.04
ECG200	95.00	96.00
ECG_five_days	81.88	92.68
Face_All	85.27	81.24
FaceFour	96.59	96.59
FacesUCR	86.34	90.29
Fish	86.86	92.00
Gun_Point	94.00	96.00
Haptics	91.23	88.31
Inline_skate	85.09	87.27
Italy_power_semand	96.89	96.89
Lighting2	95.08	100.00
Lighting7	84.93	90.41
MALLAT	90.28	96.89
Medical_images	90.26	91.45
Mote_strain	94.25	94.25
Olive_oil	96.67	96.67
OSU_leaf	83.88	87.19
Sony_AIBO_robot_surface	95.17	87.35
Sony_AIBO_robot_surfaceII	97.48	98.74
Star_light_curves	96.61	97.29
Swedh_leaf	91.04	89.28
Symbols	89.35	91.86
Synthetic_control	98.00	99.33
Trace	88.00	88.00
TwoLeadECG	90.34	90.17
Two_Patterns	92.28	95.05
uWaveGestureLibrary_X	93.30	94.89
uWaveGestureLibrary_Y	90.42	93.61
uWaveGestureLibrary_Z	91.82	94.05
Wafer	99.53	99.50
Words_synonyms	85.58	89.66
Yoga	92.30	94.07
平均值	90.88	92.61

表 5　Wilcoxon 符号秩检验

Table 5　Results of Wilcoxon signed ranks test

检验量	Accuracy		Earliness	
	统计量 z 值	概率 p 值	统计量 z 值	概率 p 值
ECTS 与 PAA_ECTS	−5.4425	5.2533e−08	−2.1705	0.0300
ECDIRE 与 PAA_ECTS	−4.0609	4.8893e−05	−3.9606	7.4767e−05
RelClass 与 PAA_ECTS	−1.6317	0.1027	−0.7124	0.4762
EDSC 与 PAA_ECTS	−5.1828	2.1856e−07	−4.7202	2.3564e−06

综上分析，本文提出的 PAA_ECTS 在保持较高可靠性的基础上明显提高了准确率，并尽可能早地实现了早期分类。

4.3　参数对分类性能的影响

本文提出的 PAA_ECTS 算法有一个参数，即 PAA 维数

d。图 1、图 2 分别给出了 PAA_ECTS 在 Gun_Point 数据集、Star_light_curves 数据集上分类的 Accuracy 随 PAA 维数的变化情况，由图 1、图 2 可以看出，随着 PAA 维数的增加，分类的准确率一直较高且趋于稳定，这说明在实际进行早期分类时 PAA 降维后每段数据量较小时即可做出分类(PAA 维数越大，每段数据量越小)。图 3、图 4 分别给出了 PAA_ECTS 在 Gun_Point 数据集、Star_light_curves 数据集上分类的 Earliness 随 PAA 维数的变化情况，由图 3、图 4 可得，在保持较高准确率的前提下，PAA_ECTS 能尽可能早地实现早期分类，即随着 PAA 参数的增加，早期性并没有明显的下降趋势。图 5、图 6 分别给出了 PAA_ECTS 在 Gun_Point 数据集、Star_light_curves 数据集上分类的 Reliability 随 PAA 维数的变化情况，由图 5、图 6 可以看出，当 PAA 维数增加时，PAA_ECTS 的可靠性基本保持稳定。

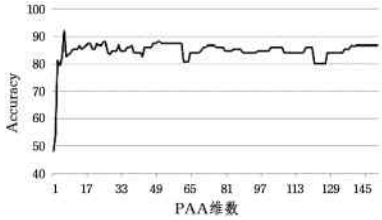

图 1 Gun_Point 数据集上 Accuracy 随 PAA 维数的变化

Fig. 1 Changing of Accuracy with the dimension of PAA on Gun_Point dataset

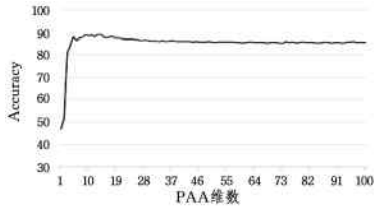

图 2 Star_light_curves 数据集上 Accuracy 随 PAA 维数的变化

Fig. 2 Changing of Accuracy with the dimension of PAA on Star_light_curves dataset

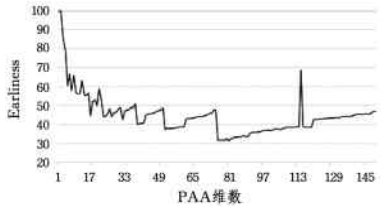

图 3 Gun_Point 数据集上 Earliness 随 PAA 维数的变化

Fig. 3 Changing of Earliness with the dimension of PAA on Gun_Point dataset

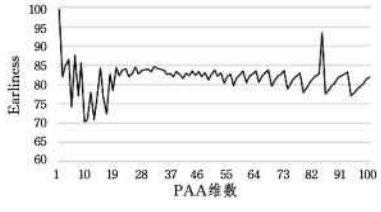

图 4 Star_light_curves 数据集上 Earliness 随 PAA 维数的变化

Fig. 4 Changing of Earliness with the dimension of PAA on Star_light_curves dataset

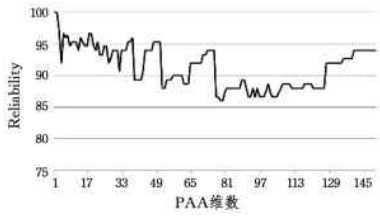

图 5 Gun_Point 数据集上 Reliability 随 PAA 维数的变化

Fig. 5 Changing of Reliability with the dimension of PAA on Gun_Point dataset

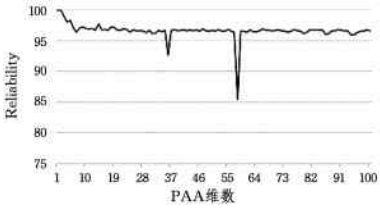

图 6 Star_light_curves 数据集上 Reliability 随 PAA 维数的变化

Fig. 6 Changing of Reliability with the dimension of PAA on Star_light_curves dataset

结束语 本文提出了基于分段聚合近似的早期分类算法 PAA_ECTS，PAA 简单直观、计算简便，针对不同的数据集，只要选择适当的 PAA 维数就可以去掉时间序列中的噪声并较好地保留时间序列的形状结构。PAA_ECTS 充分利用了 PAA 的这些优点，从而具有普适性。在不降低可靠率的情况下，PAA_ECTS 的准确率明显优于 ECTS，ECDIRE 和 EDSC，PAA_ECTS 的早期性明显优于 ECTS。如何选择“最优”的 PAA 维数 d 以及如何将 PAA_ECTS 应用于多变量时间序列的早期分类，有待今后继续研究。

参 考 文 献

[1] MA C H, WENG X Q. Review of Early Classification on Time Series[J]. Microcomputer & Its Applications, 2016, 35(16): 13-15, 19. (in Chinese)
马超红，翁小清. 时间序列早期分类综述[J]. 微型机与应用，2016，35(16)：13-15，19.

[2] PARRISH N, ANDERSON H S, GUPTA M R, et al. Classifying with confidence from incomplete information[J]. Journal of Machine Learning Research, 2013, 14(1): 3561-3589.

[3] 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.

[4] LI H L, GUO C H. Survey of feature representations and similarity measurements in time series data mining[J]. Application Research of Computer, 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.

[5] XING Z, PEI J, YU P S. Early classification on time series[J]. Knowledge & Information Systems, 2012, 31(1): 105-127.

[6] XING Z, PEI J, YU P S. Early prediction on time series: a nearest neighbor approach[C]// Proceedings of the, International Joint Conference on Artificial Intelligence (IJCAI 2009). Pasadena, California, USA, 2009: 1297-1302.

(下转第 317 页)

Processing,2007,16(5):1395-1411.

[4] XU J,SUN Y B,WEI Z H. Research on Non-Local Means Denoising Algorithm Based on Structural Tension [J]. Computer Engineering and Applications,2010,46(28):178-180. (in Chinese)
许娟,孙玉宝,韦志辉. 基于结构张量的 Non-Local Means 去噪算法研究[J]. 计算机工程与应用,2010,46(28):178-180.

[5] DABOV K,FOI A,KATKOVNIK V,et al. Image Denoising by Sparse 3D Transform-domain Collaborative Filtering [J]. IEEE Transactions on Image Processing,2007,16(8):2080-2095.

[6] HUANG M,HUANG W Q,LI J B,et al. Study on parameters based on BM3D image denoising algorithm [J]. Industrial Control Computer,2014(10):99-101. (in Chinese)
黄牧,黄文清,李俊柏,等. 基于 BM3D 图像去噪算法的参数研究[J]. 工业控制计算机,2014(10):99-101.

[7] TANG Y,TONG R,TANG M,et al. Depth incorporating with color improves salient object detection[J]. Visual Computer International Journal of Computer Graphics,2016,32(1):111-121.

[8] CHENG M M,ZHANG G X,MITRA N J,et al. Global contrast based salient region detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2011:409-416.

[9] XU D,TANG Z M,XU W. Numerical object detection of fusion color attribute and spatial information [J]. Journal of Image and Graphics,2014(4):541-548. (in Chinese)
徐丹,唐振民,徐威. 融合颜色属性和空间信息的显著性物体检测[J]. 中国图象图形学报,2014(4):541-548.

[10] MA Y F,ZHANG H J. Contrast-based image attention analysis by using fuzzy growing[C]// Proceedings of the Eleventh ACM International Conference on Multimedia. ACM,2003:374-381.

[11] ZHAI Y,SHAH M. Visual attention detection in video sequences using spatiotemporal cues [C] // ACM International Conference on Multimedia. Santa Barbara,CA,USA,DBLP,2006:815-824.

[12] GOFERMAN S,ZELNIK-MANOR L,TAL A. Context-aware saliency detection [C]// 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,2010:2376-2383.

[13] CHENG M M,ZHANG G X,MITRA N J,et al. Global contrast based salient region detection[C]// 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,2011:409-416.

[14] HAN Z,WANG H B,YU Z T,et al. Image Denoising Algorithm for Bilateral Nonlocal Mean Filter [J]. Journal of Sensors and Microsystems,2016(6):124-127,131. (in Chinese)
韩震,王红斌,余正涛,等. 双边非局部均值滤波图像去噪算法[J]. 传感器与微系统,2016(6):124-127,131.

[15] SHAO H,YU T,XU M,et al. Image region duplication detection based on circular window expansion and phase correlation [J]. Forensic Science International,2012,222(1-3):71.

[16] JAIN P,TYAGI V. An Adaptive Edge-Preserving Image Denoising Using Block-Based Singular Value Decomposition in Wavelet Domain [C]// Proceedings of the International Congress on Information and Communication Technology. 2016:19-27.

[17] CHANDLER D M. Seven Challenges in Image Quality Assessment:Past,Present,and Future Research [C]// ISRN Signal Processing. 2013:53.

[18] LI L. Empirical Study and Policy Suggestions on the Influence of Chinese Think Tanks [J]. Social Sciences,2014(4):4-21. (in Chinese)
李凌. 中国智库影响力的实证研究与政策建议[J]. 社会科学,2014(4):4-21.

(上接第 296 页)

[7] MORI U,MENDIBURU A,KEOGH E,et al. Reliable early classification of time series based on discriminating the classes over time[J]. Data Mining & Knowledge Discovery,2016,31(1):1-31.

[8] XING Z,PEI J,YU P S,et al. Extracting Interpretable Features for Early Classification on Time Series[C]// Eleventh Siam International Conference on Data Mining(SDM 2011). Mesa,Arizona,USA,2011:744-757.

[9] DING C,HE X. Cluster Aggregate Inequality and Multi-level Hierarchical Clustering[J]. Lecture Notes in Computer Science,2005,3721:71-83.

[10] CHEN Y P,KEOGH E,HU B,et al. Abdullah Mueen and Gustavo Batista[OL]. http://www.cs.ucr.edu/~eamonn/time_series_data.

[11] YUAN J D,WANG Z H. Review of Time Series Representation and Classification Techniques[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.