文章编号： 1673-3193（2024）04-0420-08

一种基于趋势距离的快速Shapelet提取算法

张苗苗，乔钢柱，李泽宇

（中北大学 计算机科学与技术学院，山西 太原 030051）

摘　要： 针对现有Shapelet提取方法无法反映趋势特点、提取结果与原始数据偏离程度略大的问题，提出了一种改进的快速Shapelet选择算法。本文首先提出了一种考虑时间序列相对趋势的距离计算方法，该方法能够更精确地度量时间序列的相似性。其次，将Shapelet特征与集成网络结合，使分类器受益于残差线性连接和注意机制，增强了算法的泛化能力。最后，在12个数据集上进行了对照试验。实验结果表明，本文方法可以获得88.0%的平均精度，与快速Shapelet算法相比平均精度提升了2.9%，尤其在ChlorineConcentration数据集上精度提高了13.3%；就加速率而言，该方法在10个数据集上的提取速度都超过了原算法，因此可以更高效地提取时间序列数据中的Shapelet。

关键词： Shapelet；趋势特征；Shapelet变换；子类划分；时间序列分类

中图分类号： TP311.13　　**文献标识码：** A　　**doi:** 10.3969/j.issn.1673-3193.2024.04.002

引用格式： 张苗苗，乔钢柱，李泽宇．一种基于趋势距离的快速Shapelet提取算法[J]．中北大学学报（自然科学版），2024，45(4)：420-427.

ZHANG Miaomiao，QIAO Gangzhu，LI Zeyu．A fast shapelet extraction algorithm based on trend distance[J]．Journal of North University of China(Natural Science Edition)，2024，45(4)：420-427.

A Fast Shapelet Extraction Algorithm Based on Trend Distance

ZHANG Miaomiao，QIAO Gangzhu，LI Zeyu

（School of Computer Science and Technology，North University of China，Taiyuan 030051，China）

Abstract： Aiming at the problem that the existing Shapelet extraction method cannot reflect the trend characteristics and the extraction result deviates slightly from the original data，an improved fast Shapelet selection algorithm was proposed. A distance calculation method considering the relative trend of time series was proposed，which could measure the similarity of time series more accurately. Secondly，the Shapelet features were combined with the ensemble network to enable the classifier to benefit from the residual linear connection and attention mechanism，which enhanced the generalization ability of the algorithm. Finally，controlled trials were conducted on 12 datasets. Experimental results show that the proposed method can obtain an average accuracy of 88.0%，which is 2.9% higher than the fast Shapelet algorithm，especially on the ChlorineConcentration dataset，and the accuracy is increased by 13.3%. In terms of acceleration rate，the method can extract faster than the original algorithm on all 10 datasets，so it can extract Shapelet in time series data more efficiently.

Key words： Shapelet；trend characteristics；Shapelet transform；subclass division；time series classification

收稿日期： 2023-09-11

基金项目： 山西省基础研究计划联合资助项目（TZLH20230818007）

作者简介： 张苗苗（1999—），女，硕士生，主要从事时间序列、大数据视觉与计算的研究。

通信作者： 乔钢柱（1999—），男，教授，博士，主要从事时间序列、工业大数据的研究。E-mail：qiaogangzhu@sohu.com。

0 引　言

在时间序列分析任务中，时间序列分类（Time Series Classification，TSC）被广泛应用。时间序列分类的目标是尽可能准确地确定未标记时间序列的类别标签，除了分类的准确性以外，分类结果的可解释性对于时间序列的相关应用也至关重要。

近年来，基于Shapelet的时间序列分类算法由于具有可解释性好、准确度高的特点受到了广泛的关注。Shapelet是时间序列数据中最显著的特征子集。例如：马鞭草和荨麻的叶片从整体来看很相似不易区分，而将它们的叶子轮廓转化为时序数据后，就会发现其主要的差别是叶柄与叶片之间的角度。因此，如果使用序列中这部分显著的特征子集（子序列）作为序列的表征，就很容易将二者区分开来。

目前的Shapelet提取方法都是通过增强Shapelet的可辨识性从而便于提取，这种方法在提取过程中通常不考虑数据中明显的趋势变化，使得提取出的Shapelet与原始数据的偏离程度变大，因而不再具有代表性。本文针对该情况提出了一种用于挖掘时间序列趋势信息的快速Shapelet选择算法网络（Trend Fast Shapelet Selection Network，TF- SSN）。该方法使用改进的欧氏距离计算方法从训练数据集中采样趋势特征明显的时间序列，然后识别其局部最远偏差点用于筛选候选Shapelet，最后将变换后的特征传入NNE（FCN-RESNET-ENCODER）中进行模型训练。与其它方法相比，本文方法在趋势特征明显的时间序列上具有更好的分类效果。

1 相关工作

1.1 TSC时间序列分类

时间序列分类方法可以分为三大类：基于模型、基于距离和基于特征的方法。

基于模型的分类方法是利用随机过程建模，不断调整模型的参数，使其达到高性能的过程。基于模型的方法多使用自回归模型、隐马尔可夫模型和深度学习模型。目前较为先进的技术正朝着集成解决的方案发展，即建立一个深度学习的模型集合。

基于距离的分类方法是定义序列之间的相似性度量，然后在具体的分类方法中以某种方式引入这些距离。这类TSC方法简单、准确、鲁棒性好，但它们的主要缺点是相似性度量不能为分类结果提供解释能力。

基于特征的分类方法是将时间序列转换为特征向量，然后使用传统的分类器对提取的特征进行分类。与全局特征相比，一些局部特征对分类结果的可解释更好。其中，基于形状的方法可解释性更加直观[1]。因此，本研究将重点放在基于形状的TSC方法上。

1.2 基于Shapelet的时间序列分类

Ye等[1]提出时序数据中Shapelet的概念后，基于Shapelet的分类器就引起了许多研究者的兴趣。目前，Shapelet发现过程中主要面临两个问题：发现过程非常耗时，无法将形状元素与其他分类器结合。

针对第一个问题，研究人员提出了一些加速Shapelet发现的策略。Yamaguchi等[2]提出Shapelet正则化方法，通过缩小适当的特征来增强特征的可辨性并保持可解释性。Rakthanmanon等[3]提出了快速Shapelet算法（Fast Shapelet，FS），基于符号聚合近似对候选对象进行了修剪。Cai等[4]提出了SS-Shapelets，利用少量标记和传播的伪标记时间序列来帮助发现具有代表性的形状。Grabocka等[5]基于目标函数学习Shapelet特征，从而可以直接学习接近最优的Shapelet。Liu等[6]提出了具有规范时间序列特征的Shapelet。Guillaume等[7]引入膨胀的概念，将Shapelet的出现次数视为一个特征。Zou等[8]提出了一种改进的基于聚类的快速Shapelet选择算法（Fast Shapelet Selection Based on Clustering，FSSoC），大大减少了Shapelet选择的时间。

针对第二个问题，一些研究者开始关注基于Shapelet的分类器。Lines等[9]将分类和Shapelet选择的过程分开。Medico等[10]在神经网络模型的体系结构中嵌入Shapelet学习，将基于Shapelet的分类扩展到多维环境。Cheng等[11]提出了新框架Time2Graph＋来学习具有时间感知能力的Shapelet。Li等[12]提出的ShapeNet能够将不同长度的候选者嵌入到统一空间中进行Shapelet选择。Ji等[13]将完全卷积网络与Shapelet特征结合从而获得了较高的精度。Yu等[14]提出了一种基于多时间尺度的Shapelet特征提取框架。詹熙等[15]提出了一种基于无监督表示学习的多变量时间序列分类方法Multi-Shapelet。

虽然已经有了许多基于Shapelet的新方法，但大多数方法都集中在提高Shapelet的鉴别性，并未

对原始时间序列的趋势信息加以利用，使得Shapelet发现过程还是较为耗时且在特征增强过程中使得Shapelet偏离真实子序列过多、泛化性能较差，这与Shapelet方法的初衷相悖。另一种主流方法则是在深度学习过程中直接学习Shapelet，这类方法集中在通过各种方法提高分类器的性能，在学习Shapelet的过程中很少添加约束，即学习到的Shapelet应该类似于子序列，使得Shapelet可解释性降低。

1.3 时间序列的趋势性

趋势性是时间序列的一种重要特性，它能够直观地反映时间序列的变化趋势，即在一定时间段内的单调性。一些研究者开始发现并利用时间序列的这种特性。李宏伟等[16]提出了一种自动提取时间序列趋势转折点的算法，为定量化提取地震异常信息提供了客观的基础信息。Zhao等[17]提出了动态多视角个性化相似度测度用于度量股票价格曲线之间的相似程度。刘意杨等[18]提出了基于转折点和趋势段的时间序列趋势提取算法。杜加础等[19]提出了一种基于模态分量重构及多维评价的时间序列趋势提取算法。这些趋势信息提取的算法在提取过程中不能准确提取转折点导致拟合误差较大、难以反映时间序列的整体趋势，目前将这种趋势性应用于基于Shapelet的方法还比较少。

2 相关定义

本文涉及的相关符号如表1所示。

表1 本文涉及的相关符号

Tab. 1 Related symbols covered in this article

符号	说明	符号	说明
D	数据集	*SubDist*	子序列与时间序列的距离
T	时间序列	*IG*	信息增益
S	子序列	I	熵
m	时间序列长度	*dth*	距离阈值
L	子序列长度	*Gain*	增益
Dist	序列间的距离	*LFDP*	局部最远偏差点

定义1：单变量时间序列。单变量时间序列$T=\{t_1,t_2,t_3,\cdots,t_L\}$是$L$个实值变量的有序集合。数据点$t_1,t_2,t_3,\cdots,t_L$通常按时间顺序排列，间隔时间相等，$t_i$是时间戳$i$处的值。

定义2：时间序列的子序列。给定长度为L的时间序列T，子序列S是在T的连续位置上得到的长度为l的样本，即$S=\{t_p,\cdots,t_{p+l-1}\}$。

定义3：时间序列之间的距离。给定长度为L的两个时间序列T和R，T与R之间的距离可用$Dist(T,R)$表示。以欧氏距离为测度，$Dist(T,R)$的计算公式为

$$Dist(T,R)=\sqrt{\frac{1}{L}\sum_{i=1}^{L}(t_i-r_i)^2}。\tag{1}$$

定义4：时间序列与子序列之间的距离。距离函数$SubDist(T,S)$以时间序列T和子序列S作为输入，返回一个非负值$d(d\geqslant 0)$，即T到S的距离，函数公式为

$$SubDist(T,S)=\min(Dist(S,S')),S'\in S^l(T)。\tag{2}$$

定义5：熵。一个时间序列数据集D由两个子类的数据组成，分别标记为A和B。假设A类中时间序列对象的比例为$p(A)$，B类中时间序列对象的比例为$p(B)$，则D的熵为

$$I(D)=-p(A)\log(p(A))-p(B)\log(p(B))。\tag{3}$$

每个分割策略将整个数据集D分成两个子集，D_1和D_2。因此，分割后整个数据集中剩余的信息由每个子集的加权平均熵来定义。D_1中物体的比例为$f(D_1)$，D_2中物体的比例为$f(D_2)$。所以，D分裂后的总熵是

$$\hat{I}(D)=f(D_1)+I(D_1)+f(D_2)+I(D_2)。\tag{4}$$

定义6：最佳分割点(Optimal Split Point，OSP)。一个时间序列数据集D由A和B两类组成。对Shapelet候选者S，选择一些距离阈值dth，并将D分为D_1和D_2，这样对于D_1中的每个时间序列对象T，$SubDist(T,S)<dth$，对于D_2中的每个时间序列对象T，$SubDist(T,S)>dth$。

定义7：Shapelet。在一个由两种类别组成的时间序列数据集D中，$shapelet(D)$是其中一个子序列，对于任何子序列S，与之相对应的最佳分割点为

$$Gain(Shapelet,\ d_{OSP(D,Shapelet(D))}),$$
$$Gain(Q,\ d_{OSP(D,S)})。\tag{5}$$

由于Shapelet只是任何长度小于或等于数据集中最短时间序列长度的时间序列，因此它可以拥有无限多可能的形状。一个类中的时间序列对象可能包含一些类似的子序列，这些子序列被视为Shapelet的候选对象。

定义8：局部最远偏差点(Local Farthest Deviation Point，LFDP)。该点在S序列中权值最大，且与S的拟合线距离最大。一个子序列的权值表示为$weight=\max(dist_{\text{sum}},\ 2*dist_{\max})$，$dist_{\text{sum}}$是子序列中所有点的距离之和，$dist_{\max}$是这些距离

中的最大值。距离是点对拟合线的拟合误差。对按权重设置的时间序列片段进行排序，选择权重最大的时间序列片段，LFDP 即为所选时间序列段的拟合线距离最大的点。

定义 9：Shapelet 变换（Shapelet Transform，ST）。Shapelet 变换将时间序列转换为新的特征空间。将 Shapelet 视为一个特征，Shapelet 与时间序列之间的距离对应于该特征的值。假设选择的 Shapelet 集合为 $\{Shapelet\}$。变换后的时间序列为

$$T_{\text{transformed}} = SubDist(Shapelet_1, T), SubDist(Shapelet_2, T), \cdots, SubDist(Shapelet_{|Shapelet|}, T), \quad (6)$$

式中：$|Shapelet|$ 为 $\{Shapelet\}$ 中 Shapelet 的个数。

3　TFSSN 方法

本文提出的方法 TFSSN 能够在不损失特征可解释性的前提下有效地从时间序列中提取趋势信息，使得提取到的 Shapelet 特征更具有鉴别性，从而达到较高的 TSC 精度。TFSSN 分为 3 个阶段：

1）Shapelet 特征提取，提取判别时间序列子序列作为 Shapelet 特征；

2）Shapelet 转换，采用基于趋势的欧氏距离计算方法计算得出变换后的 Shapelet 特征向量；

3）分类器训练，构造一个 NNE 作为分类器，以达到较高的准确率。

TFSSN 的整体流程如图 1 所示。

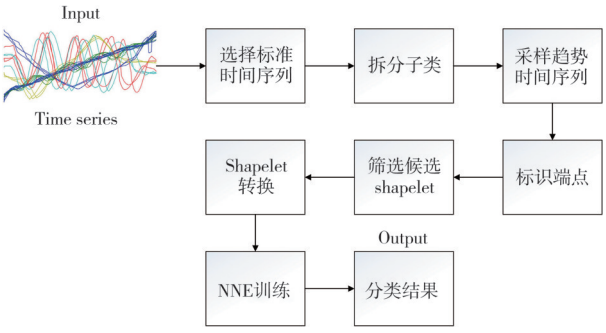

图 1　TFSSN 整体流程

Fig. 1　The overall process of TFSSN

3.1　Shapelet 特征提取

特征提取过程包含选择标准时间序列、拆分子类、采样时间序列、筛选候选 Shapelet 四个步骤。

1）选择标准时间序列，即选择最接近和均值的时间序列作为标准时间序列。假设数据 T 为 k 个长度为 l 的同类时间序列的集合，表示为

$$T = \{T_1, T_2, T_3, \cdots, T_n, \cdots, T_k\}, \quad (7)$$

$$T_n = \{T_{n_1}, T_{n_2}, T_{n_3}, \cdots, T_{n_m}, \cdots, T_{n_l}\}。 \quad (8)$$

将每个序列中的所有值相加，以计算这类特定数据的平均值，如式（9）与式（10）所示。

$$S_n = \sum_{i=1}^{l} T_{n_i}, \quad (9)$$

$$Mean = \left(\sum_{i=1}^{k} S_i\right)/k。 \quad (10)$$

式（9）中最接近均值的序列将被选为标准时间序列。

$$PivotP = argmin_{n\in\{1,\cdots,k\}}(|Mean - S_n|)。 \quad (11)$$

2）拆分子类。在典型的训练数据中，同一类中的序列可能具有不同的特征，称为子类。先发现数据每个类中的子类，再从中进行采样，这个过程可以快速地为每个类别采样合适数量的时间序列，减少参数冗余，提高候选 Shapelet 的质量。为了加速这一过程和更好地体现两个序列之间的差异[20]，本文使用欧式距离计算方法计算类中所有其他数据序列与准则之间的距离并进行排序，得到相邻距离值的差值并计算其标准差值。最后，按照拆分差异大于计算标准差一半的序列来将数据分成子类。

3）采样时间序列，即从每个子类中采样一个时间序列。在每个子类中，选取与其他时间序列相比趋势距离和值最小的时间序列作为样本。具体过程见 3.2。

4）筛选候选 Shapelet。使用基于重要数据点的时间序列数据分段线性表示方法来获得时间序列的端点。包含特殊点的子序列和以非相邻 LFDP 为端点的子序列具有更高的判别性，所以将被选为 Shapelet 候选者。

3.2　Shapelet 转换

通过将时间序列实例转换为 Shapelet 特征向量，充分利用了 Shapelet 特征的可解释性，如式（12）所定义。

$$T' = \{dist(T, S_1), \cdots, dist(T, S_i), \cdots, dist(T, S_k)\}。 \quad (12)$$

大多数算法使用欧氏距离来度量相似性，但欧氏距离只是计算点到点的距离，而不考虑时间序列的变化趋势。本文认为如果两个时间序列呈现相似的趋势，并且它们之间的距离很小，则这两个时间序列是相似的。在传统欧式距离计算中加入趋势信息的考量能够更精确、更快速地度量时间序列的相

似性。据此，本文提出了一种基于趋势的欧氏距离计算方法，两个序列的相似度计算公式为

$$TDist(T,R)=\sqrt{\frac{1}{L}\left(\sum_{i=I_s}(t_i-r_i)^2+\sum_{j=I_s}(t_j-r_j)^2*\lambda\right)},\quad(13)$$

式中：T和R为长度相同的两个时间序列；t_i,r_i,t_j,r_j表示时间序列在时间戳i或j处的值；I_s是I的反数；$\lambda\in[1,1.5]$。

在具体的计算过程中，判断两个序列是否具有相似的趋势，就要去查看两个序列的每个时间戳处的两个点之间的大小关系是否相同。若相同，则直接将时间戳处的距离进行叠加，否则，应用趋势参数λ小范围地扩大距离。因此，趋势相似的时间序列距离会更小。在图2的例子中，观察变化趋势可以发现A与B更相似，但是直接使用欧式距离计算A与B、C之间的距离，得出的结果是一致的，使用基于趋势的欧氏距离公式计算得出的结果则是A与B更相似。这表明，基于趋势的欧氏距离计算方法可以更好地衡量时间序列之间的相似性。

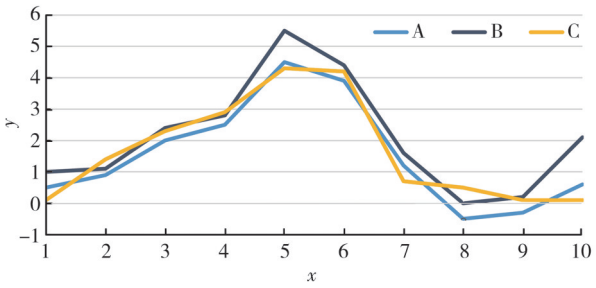

图2 时间序列A、B、C之间的趋势关系

Fig. 2 The trend relationship between time series A, B, and C

3.3 分类器训练

尽管一些研究人员使用深度学习获得了Shapelet特征，但很少有人采用深度学习方法对Shapelet转换后的数据集进行分类。TFSSN的最后一步是为变换后的向量训练NNE分类器。NNE是仅由ResNet、FCN和编码器组成的神经网络集成。将FCN与ResNet和Encoder相结合，可以使分类器分别受益于残差线性连接和注意机制。

将Shapelet转换后的特征向量依次传入ResNet、FCN和编码器进行训练并输出相应的独热编码分类结果，NNE对3个结果进行加权平均，保存集成后准确率最高的权重分配结果。权重的具体分配由具体的函数实现，设置权重总和为30，依次遍历计算每种情况下的分类准确率。权重分配的结果基本符合一个观念，即为准确率高的分类器分配更高的权重。

4 实验部分

4.1 实验设置

本文从UEA & UCR时间序列分类存储库中选择了12个数据集[21]进行实验，这12个数据集涵盖了如医学图像、电力需求等广泛领域，且在类别数量、训练集大小和时间序列长度方面覆盖范围很广。本文在实验中使用训练数据来构建TSC方法，使用测试数据来验证分类精度。数据集D的准确率为

$$Accuracy=\frac{N_c}{N_{test}}。\quad(14)$$

式中：N_{test}为测试数据点的总数；N_c为相应测试上正确分类的个数。

在所有的实验中，选择的Shapelet数量被设置为训练集大小的一半，这个参数是根据经验设置。LFDP的个数设置为$[0.05*m+2]$，其中，m为时间序列长度，2表示必须选择时间序列中的第一个点和最后一个点作为LFDP。在趋势距离计算中，本文使用每个数据集的训练集进行交叉验证，得到λ的值，λ的范围设置在$[1,1.5]$之间。对于网络训练部分，集成网络权重设置采用函数实现，能够自动选取组合最优的结果，网络训练学习率针对不同数据集采用不同的学习率，各数据集的参数设置如表2所示。

表2 参数设置

Tab. 2 Parameter setting

数据集	参数λ	参数l_r
Adiac	1.031 9	1
Beef	1.031 9	1
ChlorineConcentration	1.3	1
Coffee	1.031 9	1
DiatomSizeReduction	1.1	0.001
ItalyPowerDemand	1.1	0.000 1
Lighting7	1.2	0.000 1
MedicalImages	1.3	1
MoteStrain	1.1	0.01
Symbols	1.5	0.001
Trace	1.031 9	0.000 1
TwoLeadECG	1.1	0.000 1

4.2 对比试验

4.2.1 TFSSN分类精度对比实验

为了展示TFSSN的整体性能，本文将TFSS与TFSSN的分类结果共同进行展示。尤其为了验证TFSS这一数据转换方法的有效性，本文对比了

TFSS、3种经典的基于Shapelet的方法（FS、ST、LS）和2种基于快速Shapelet选择的算法（FSS[22]、FSSoC[8]）。6种基于Shapelet的算法的分类准确率如表3所示，表中还给出了平均精度和平均精度等级。

表3　TFSS分类准确率对比

Tab. 3　Comparison of TFSS classification accuracy

数据集	ST	LS	FS	FSS	FSSoC	TFSS
Adiac	0.783	0.522	0.593	0.780	0.769	**0.816**
Beef	**0.900**	0.867	0.567	0.833	0.833	0.833
ChlorineConcentration	0.700	0.592	0.546	0.607	0.655	**0.740**
Coffee	0.964	**1.000**	0.929	0.929	**1.000**	**1.000**
DiatomSizeReduction	0.925	**0.980**	0.866	0.912	0.898	0.876
ItalyPowerDemand	0.948	0.960	0.917	0.926	0.956	**0.965**
Lighting7	0.726	**0.795**	0.644	0.740	0.767	0.753
Medical	0.670	0.664	0.624	0.721	0.710	**0.736**
MoteStrain	0.897	0.883	0.777	0.899	0.895	**0.900**
Symbols	0.882	0.932	**0.934**	0.895	0.913	0.913
Trace	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
TwoLeadECG	0.997	0.996	0.924	0.972	0.991	**0.998**
Average acc	0.866	0.849	0.777	0.851	0.866	**0.878**
Average rank	3.00	2.83	5.00	3.50	2.92	1.83

由对比实验结果可以发现，TFSS的平均精度是最高的。这表明TFSS在时间序列分类中比其它方法的加速策略更加稳定，对趋势明显的数据提取更有效。增加了趋势参数λ后，两个序列间的相似度得以动态调整，不再单单考虑距离的大小，而是将序列的变化趋势也纳入考量。通过加深趋势变化在序列相似性中发挥的作用，大大提升了TFSS分类的能力。

表4　NNE分类准确率对比

Tab. 4　Comparison of NNE classification accuracy

数据集	ResNet	FCN	Encoder	LSTM	NNE
Adiac	**0.737**	0.645	0.678	0.678	**0.737**
Beef	0.700	0.800	0.767	**0.833**	**0.833**
ChlorineConcentration	0.647	0.663	0.610	0.612	**0.675**
Coffee	0.964	0.964	**1.000**	**1.000**	**1.000**
DiatomSizeReduction	0.925	0.438	**0.974**	**0.974**	**0.974**
ItalyPowerDemand	0.966	**0.970**	0.962	0.966	**0.970**
Lighting7	0.671	0.685	0.712	0.767	**0.795**
Medical	0.728	0.733	0.718	0.728	**0.758**
MoteStrain	0.895	0.891	0.893	**0.909**	0.903
Symbols	0.938	**0.943**	0.938	0.936	**0.943**
Trace	0.990	0.990	**1.000**	**1.000**	**1.000**
TwoLeadECG	0.958	0.951	0.949	0.963	**0.969**

表4中对比了ResNet、FCN、Encoder、LSTM和NNE在各数据集上的分类准确率。不难发现，将获取的Shapelet与NNE结合后，所获得的分类精度整体上是最高的。在时间序列分类时选择合适的分类器进行集成会产生更好的融合效果和更高的准确率[23]。例如，在DiatomSizeReduction数据集上，单独使用FCN很容易产生过拟合，而通过NNE集成后却能实现97.4%的准确率，这意味着ResNet、FCN和Encoder的组合使分类器能够分别受益于残差线性连接和注意力机制，在其他数据集上的结果也表明这一组合具有很强的泛化能力。

4.2.2　Shapelet提取时间比较

本文采用加速率（r）来评价加速效果，加速率（r）是ST的Shapelet提取时间（t_{ST}）与其他相应方法的Shapelet提取时间（t_{cor}）之比，如式（15）所示。r值越大，表示加速度效应越大。

$$r=\frac{t_{ST}}{t_{cor}}。\tag{15}$$

由于本文算法是在FSS算法基础上进行的改进，为了体现TFSS能更有效地提取Shapelet，除准确率外，本文对两种算法的提取时间也进行了比较。如表5所示，TFSS提取Shapelet特征的速度在大多数数据集上都比原始的FSS方法快。这是由于TFSS在提取过程中针对时间序列具体时间轴处的距离进行了动态调整，选择了更相似的Shapelet，从而达到了高精度和低耗时。而在MoteStrain等数据集上提升速度不明显的原因主要在于这类数据集原始序列长度较短，本文方法基于趋势性进行Shapelet提取的优势无法体现，因而无法有效地进行加速。但大多数时间序列都比较长，所以可以认为本文方法在提取时间上是有一定优势的。

表5　加速率比较

Tab. 5　Comparison of speedup rates

数据集	ST	FSS	TFSS
Adiac	1	520.09	**521.07**
Beef	1	**298.35**	275.15
ChlorineConcentration	1	1679.13	**1798.61**
Coffee	1	599.90	**685.32**
DiatomSizeReduction	1	476.50	**541.19**
ItalyPowerDemand	1	294.58	**458.07**
Lighting7	1	1348.07	**1453.33**
MedicalImages	1	2920.51	**2933.96**
MoteStrain	1	**172.00**	149.09
Symbols	1	614.21	**616.64**
Trace	1	638.11	**719.45**
TwoLeadECG	1	136.44	**161.63**

4.2.3　参数λ的设置

为了测试参数λ对分类准确率的影响，本文将λ分别设置为1.0319，1.1，1.2，1.3，1.4，1.5，1.6，1.7，1.8和1.9，同时在保证其他参数不变的情况下对每个数据集进行试验，在此选取3个数据集（Lighting7、MedicalImages和ChlorineConcentration）进行展示，图3展示了不同λ对实验结果准确率的影响。由图3可以看出，在1.5

之后，随着λ增大，3个数据集的准确率虽然部分区域有小幅度的上升，但整体趋势是下降的，这意味着参数设置过大会导致提取到的序列严重偏离原始序列，进而使得分类准确率降低。根据3个数据集准确率峰值的不同可以得出，参数的设置也不是一定的，应根据不同的数据集设置相应的λ。

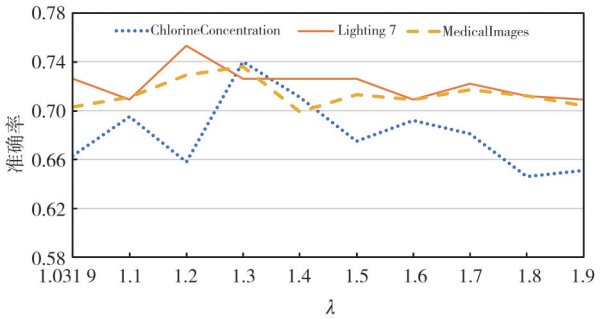

图3 参数λ对分类准确率的影响

Fig. 3 Effect of parameter λ on classification accuracy

4.2.4 心电信号特征提取案例

本节以ECG200数据集为例，通过趋势特征提取方法演示选定的Shapelet。心电图(Electrocardiogram，ECG)信号记录的是心脏在多个心动周期中所产生的生物电信号。图4展示的是一个完整的心电图波形，也称为P-QRS-T波，进行特殊标注的部分是ST波。正常心电图与心肌梗死心电图的区别主要体现在ST波的变化上。

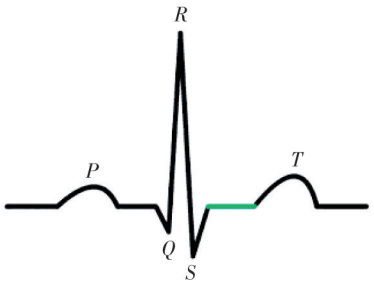

图4 一个心跳波

Fig. 4 A heartbeat wave

如图5所示，提取出的Shapelet基本对应于ST波，没有明显的上下偏移，且具有平缓的趋势，表明心跳正常。图6所提取的Shapelet也包含ST波，可以看到，ST波显著升高，形成弓背样变化，甚至有T波倒置的趋势，这是心肌梗死急性期的表现。从以上的分析可以得出，本文提取出的Shapelet可以通过不同类别的趋势特点反映两个类别之间的差异，并能达到较高的准确性。

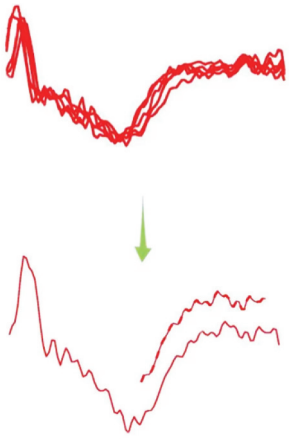

图5 正常心跳波

Fig. 5 Normal heartbeat wave

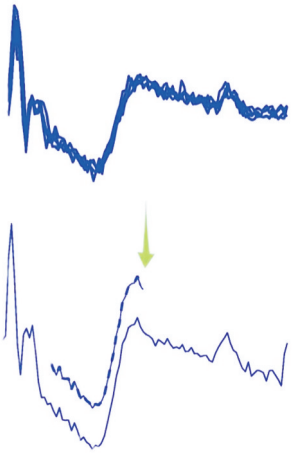

图6 异常心跳波

Fig. 6 Abnormal heartbeat waves

5 结 论

本文针对Shapelet提取的准确率不高、速度较慢且在深度学习方面应用不广泛等问题，引入了TFSSN提取算法，设计了考虑趋势信息的欧式距离计算方法，改进了原始NNE中模型权重的分配过程，充分利用了原始时间序列中的数据信息和深度学习模型。在UEA & UCR时间序列分类库的多个数据集上的实验结果表明，该方法能够更准确、有效地提取出Shapelet特征，尤其在趋势信息明显的数据上效果显著。但由于算法在Shapelet发现过程中着重保留了时间序列的趋势信息，在对部分趋势信息较弱的时间序列进行分类时，分类效果会差于趋势特征明显的数据，因此在这方面还有待提高。

参考文献：

[1] YE L, KEOGH E. Time series Shapelets: a novel

technique that allows accurate, interpretable and fast classification[J]. Data Mining and Knowledge Discovery, 2011, 22: 149-182.

[2] YAMAGUCHI A, UENO K, KASHIMA H. Learning time-series Shapelets enhancing discriminability [C]//2022 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics, 2022: 190-198.

[3] RAKTHANMANON T, KEOGH E. Fast Shapelets: A scalable algorithm for discovering time series Shapelets [C]//2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2013: 668-676.

[4] CAI B R, HUANG G Y, YANG S Q, et al. SS-shapelets: Semi-supervised clustering of time series using representative shapelets[DB/OL]. (2023-04-18)[2023-09-11]. http://arxiv.org/pdf/2304.03292v2.

[5] GRABOCKA J, SCHILLING N, WISTUBA M, et al. Learning time-series Shapelets[C]//20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014: 392-401.

[6] LIU H Y, GAO Z Z, WANG Z H, et al. Time series classification with Shapelet and canonical features[J]. Applied Sciences, 2022, 12(17): 8685.

[7] GUILLAUME A, VRAIN C, ELLOUMI W. Random dilated Shapelet transform: A new approach for time series Shapelets[C]//International Conference on Pattern Recognition and Artificial Intelligence. Cham: Springer International Publishing, 2022: 653-664.

[8] ZOU X, ZHENG X, JI C, et al. An improved fast Shapelet selection algorithm and its application to pervasive EEG[J]. Personal and Ubiquitous Computing, 2022, 26: 941-953.

[9] LINES J, DAVIS L M, HILLS J, et al. A Shapelet transform for time series classification[C]//18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012: 289-297.

[10] MEDICO R, RUYSSINCK J, DESCHRIJVER D, et al. Learning multivariate Shapelets with multi-layer neural networks for interpretable time-series classification[J]. Advances in Data Analysis and Classification, 2021, 15(4): 911-936.

[11] CHENG Z, YANG Y, JIANG S, et al. Time2Graph+: Bridging time series and graph representation learning via multiple attentions[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(2): 2078-2090.

[12] LI G Z, CHOI B, XU J L, et al. Shapenet: A Shapelet-neural network approach for multivariate time series classification[C]//AAAI Conference on Artificial Intelligence, 2021: 8375-8383.

[13] JI C, HU Y, LIU S, et al. Fully convolutional networks with Shapelet features for time series classification[J]. Information Sciences, 2022, 612: 835-847.

[14] YU H, XU C, GENG G, et al. Multi-time-scale shapelet based feature extraction for non-intrusive load monitoring [J]. IEEE Transactions on Smart Grid, 2024, 15(1): 1116-1128.

[15] 詹熙, 黎维, 潘志松. Multi-shapelet: 一种基于shapelet的多变量时间序列分类方法[J]. 数据采集与处理, 2023,38(2): 386-400.
ZHAN Xi, LI Wei, PAN Zhisong. Multi-shapelet: A multivariate time series classification method based on shapelet[J]. Journal of Data Acquisition and Processing, 2023,38(2): 386-400. (in Chinese)

[16] 李宏伟, 闫伟. 时间序列的趋势转折点提取算法及应用研究[J]. 大地测量与地球动力学, 2020, 40(12): 1242-1247.
LI Hongwei, YAN Wei. Research and application of trend turning point extraction algorithm for time series data[J]. Journal of Geodesy and Geodynamics 2020, 40(12): 1242-1247. (in Chinese)

[17] ZHAO F, GAO Y, LI X, et al. A similarity measurement for time series and its application to the stock market [J]. Expert Systems with Applications, 2021, 182: 115217.

[18] 刘意杨, 李俊朋, 白洪飞, 等. 基于转折点和趋势段的时间序列趋势特征提取[J]. 计算机应用, 2020, 40(S1): 92-97.
LIU Yiyang, LI Junpeng, BAI Hongfei, et al. Trend feature extraction method for time series based on turning point and trend segment[J]. Journal of Computer Applications,2020,40(S1): 92-97. (in Chinese)

[19] 杜加础, 车文刚, 程文辉. 基于模态重构与多维评价的时间序列趋势提取[J]. 重庆邮电大学学报(自然科学版), 2022, 34(5): 902-913.
DU Jiachu, CHE Wengang, CHENG Wenhui. Trend feature extraction method for time series based on mode reconstruction and multidimensional evaluation [J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2022, 34(5): 902-913. (in Chinese)

[20] IGLESIAS F, KASTNER W. Analysis of similarity measures in times series clustering for the discovery of building energy patterns[J]. Energies, 2013, 6(2): 579-597.

（下转第 480 页）