ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 2,Mar. 2023,pp. 386－400
DOI:10. 16337/j. 1004-9037. 2023. 02. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

Multi-shapelet:一种基于shapelet的多变量时间序列分类方法

詹　熙，黎　维，潘志松

（中国人民解放军陆军工程大学指挥控制工程学院，南京 210007）

摘　要： shapelet是时间序列中最具有辨识性的子序列，其一经提出就被来自各个领域的研究人员广泛研究，并在此过程中提出了许多有效的shapelet发现技术用于进行时间序列分类。然而，多变量时间序列的候选shapelet可能长度不同且变量来源不同，故很难直接对其进行比较，这对基于shapelet多变量时间序列分类方法提出了独特的挑战。为了应对这一挑战，提出了一种基于无监督表示学习和shapelet的多变量时间序列分类方法Multi-shapelet。Multi-shapelet首先使用混合模型DC-GNN（Dilated convolution neural network and graph neural network，DC-GNN）作为编码器，将不同长度的候选shapelet嵌入统一的shapelet选择空间，以进行shapelet之间的比较；其次，提出了一种新的损失函数以无监督学习方式训练该编码器，使得DC-GNN对shapelet编码得到相应的嵌入（Embedding）后，属于同类shapelet对应的嵌入之间的相对位置形成的拓扑与原空间中shapelet之间相对位置形成的拓扑之间的关系更接近于一种等比例的缩小，这对后续基于相似性的剪枝过程十分重要；最后，使用K-means聚类和模拟退火算法进行shapelet剪枝和选择操作。在UEA的18个多变量时间序列数据集上的实验结果表明，Multi-shapelet的整体精度相比于其他方法得到了显著提升。

关键词： shapelet；无监督表示学习；K-means聚类；模拟退火算法；shapelet剪枝

中图分类号： TP181;O211.61　　**文献标志码：** A

Multi-shapelet : A Multivariate Time Series Classification Method Based on Shapelet

ZHAN Xi，LI Wei，PAN Zhisong

（College of Command and Control Engineering，The Army Engineering University of PLA，Nanjing 210007，China）

Abstract： Shapelet is the most identifiable subsequence in time series，which has been extensively studied by researchers from various fields since it was proposed. In this process，many effective shapelet discovery techniques have been proposed for time series classification. However，candidate shapelets of multivariate time series may have different lengths and different sources of variables，making it difficult to directly compare them，which presents a unique challenge to the classification method of multivariable time series based on shapelet. we propose Multi-shapelet，a multivariate time series classification method based on unsupervised representation learning and shapelets. Firstly，Multi-shapelet uses a hybrid model DC-GNN（Dilated convolution neural network and graph neural network）as an encoder to embed candidate shapelets of different lengths into a unified shapelet selection space for comparison between shapelets. Secondly，a

基金项目：国家自然科学基金(62076251)。
收稿日期：2022-05-18；**修订日期：**2022-09-05

new loss function is proposed to train the encoder in an unsupervised learning manner, so that after DC-GNN encodes the shapelet to obtain the corresponding embedding, the topology and the original space formed by the relative positions between the embeddings corresponding to the shapelet belonging to the same class. The relationship between the topologies formed by the relative positions of the shapelet in the middle is closer to a proportional reduction, which is very important for the subsequent similarity-based pruning process. Finally, the K-means clustering and simulated annealing algorithm are proposed to prune and select shapelets to select a set of shapelets with strong classification ability. Experimental results on 18 UEA multivariable time series datasets show that the overall accuracy of Multi-shapelet is significantly better than other methods.

Key words: shapelet; unsupervised representation learning; K-means clustering; simulated annealing algorithm; shapelet prune

引　言

shapelet是文献[1]提出的一个概念,表示时间序列中能够最大程度区分时间序列类别的子序列。由于shapelet在分类问题中具备良好的可解释性和较高的精确度,近年基于shapelet的时间序列分类方法的研究取得了很大的进展。然而,这些进展大多局限于单变量时间序列分类问题[2-4],因为在多变量时间序列分类问题中,同一个数据集中的候选shapelet可能具有不同的长度且变量来源不同,这导致很难对其进行比较。

针对上述问题,研究者们提出了不同的解决方案。Ma等[5]提出了一种基于对抗生成网络的shapelet学习方法,通过训练对抗生成网络模型,使其具备生成相应的shapelet进行分类的能力,该方法虽然有效提高了模型的分类精度,但由于其所学习到的shapelet均为模型生成而不是直接从序列中提取的,因此对于shapelet的可解释性具有一定的损害。Li等[6]提出了一种先编码后剪枝的多变量时间序列分类方法ShapeNet,通过训练一个编码器将候选shapelet映射到维度统一的空间,得到不同候选shapelet长度统一的嵌入,从而消除由于候选shapelet长度不同和变量来源不同所导致的难以直接比较的问题,并提出了一种基于簇的三元组损失函数以无监督的方式训练该编码器;接着,在此基础上,提出使用基于K-means聚类的剪枝方法和评价函数进行shapelet的剪枝和筛选;最后,在选出的shapelets集合的基础上,使用多变量shapelet转换技术[7]和支持向量机对数据集进行分类。虽然该方法的提出有效解决了候选shapelet难以比较的问题,但也带来了其他问题:(1)使用编码器学习候选shapelet长度统一的嵌入时,花费的时间太多,影响了该方法的效率;(2)其所提的损失函数在训练模型时会导致如图1所示的情况出现,在映射前,原空间中样本6/7是离正/负样本簇中心最近的点,而映射后,新空间中离正/负样本簇中心最近的点却变成了样本点5/2对应的嵌入。这种情

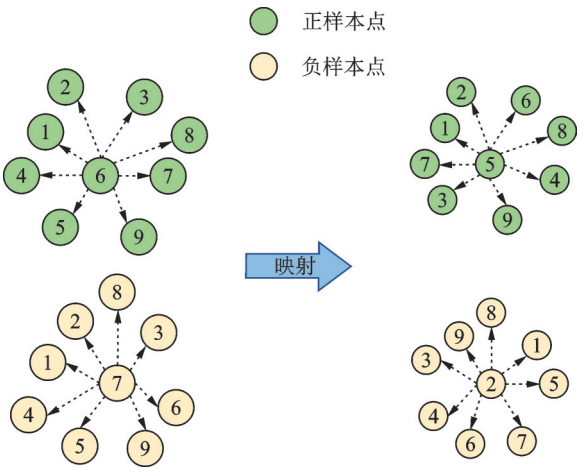


图1　在使用三元损失函数情况下正/负样本点在映射前后的相对位置拓扑图

Fig.1　Relative position topology of positive/negative sample points before and after mapping using triple loss function

况的发生将会对后续基于相似性的剪枝过程造成负面影响，从而导致模型在训练的过程中损失值不断降低而最终分类精度不升反降；(3)所提的评价函数在面对不同数据集的shapelet进行筛选时不够灵活，较难选出分类能力强的shapelet集合，而这将直接影响最终的分类精度。

以上问题的存在制约着模型最终的分类精度和效率，因此，为了提高模型的分类精度和效率，本文分别在编码器、损失函数以及shapelet选择方法这3个方面做出了相应的改进。本文的主要贡献如下：

(1) 提出使用基于扩张因果卷积网络和图神经网络的混合模型DC-GNN作为编码器，利用扩张因果卷积提取时间序列多尺度特征的能力和图神经网络在社群发现问题上速度较快的优点，加快在原空间中同类样本点对应嵌入的聚集速度，减少模型的训练时间。

(2) 提出一个新的损失函数，其可以使正/负样本对应的嵌入之间相对位置形成的拓扑与原空间中样本点之间相对位置形成的拓扑之间的关系更接近于一种等比例的缩小(相关原理在2.2.2节进行了详细阐述)，从而使得在映射前后，正/负样本在原空间和新空间相对位置的变化更类似于图2所示，保证了样本点从原空间映射到新空间后不会对后续基于聚类的剪枝造成不良影响。

(3) 提出使用基于模拟退火算法的特征选择方法筛选最终的K个shapelet集合。在使用基于K-means聚类进行shapelet剪枝后，为了筛选出分类能力强的shapelet集合，相比于使用固定的评价函数来筛选最终的K个shapelet，本文提出的基于模拟退火的shapelet选择方法更灵活，更容易获得分类能力强的shapelet集合。

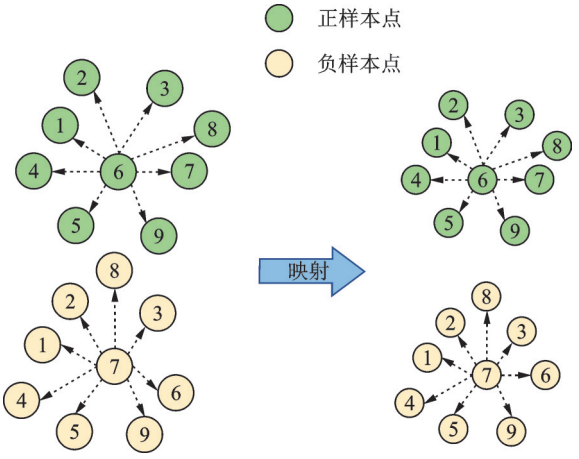


图2 在使用本文所提损失函数情况下正/负样本点在映射前后的相对位置拓扑图

Fig.2 Relative position topology of positive/negative sample points before and after mapping using the proposed loss function

1 相关工作

1.1 基于数据驱动的多变量时间序列分类方法

基于数据驱动的多变量时间序列分类方法主要以基于深度学习的方法为主。深度学习方法通过数据驱动实现分类，利用深层神经网络的高度非线性计算单元自动拟合数据自身的潜在规律，从而实现对多变量时间序列数据分类的目的。Zou等[8]提出了基于残差连接的卷积神经网络ResCNN，该网络利用基于残差连接的卷积神经网络(Convolutional neural network, CNN)提取特征进行分类。Karim等[9]提出了MLSTM-FCNs模型用于多变量时间序列分类，该模型采用长短时记忆神经网络(Long short term memory, LSTM)层和堆叠的CNN层来提取时间序列的特征，最后根据提取的特征用归一化指数函数(Softmax)层进行分类。与上述方法类似的是TST[10]和TabTransformer[11]方法，它们利用Transformer等模块提取时序特征，然后进行分类。以上基于深度学习的分类方法虽然在分类精度上与之前方法相比具有显著的提升，但在其可解释性上无法满足需要。

1.2 基于特征驱动的多变量时间序列分类方法

基于特征驱动的方法可以细分为基于相似性特征的方法和基于结构特征的方法。基于相似性特征的方法主要利用原始序列之间的相似性度量进行分类，最常用的距离度量计算指标为欧式距离(Euclidean distance，ED)[12]和动态时间规整(Dynamic time warping，DTW)[13]。上述距离度量最初用于单

变量时间序列分类任务,Shokoohi等[14]将ED和DTW推广至多变量时间序列的场景中,提出了将独立欧氏距离(ED independent, EDI)和独立动态时间规整(DTW independent, DTWI)等方法用于多变量时间序列分类。

基于结构特征的方法[15-16]通过提取原始时间序列的统计信息或转换原始时间序列的表示进行分类,这些统计信息包括数据样本的最大值、最小值、平均值、方差、频率、周期和shapelet等信息。Franceschi等[17]提出了负样本(Negative samples, NS)方法,该方法先通过一种无监督方式训练编码器,然后学习时间序列样本的低维特征表示,最后使用SVM对该特征表示进行分类。Li等[6]在该NS方法思想基础上提出了一种基于shapelet的多变量时间序列分类方法ShapeNet,通过对候选shapelet先编码后筛选,从而筛选出最终的shapelet,实验结果表明,该方法所提取的shapelet具备良好的可解释性和精度。

2　Multi-shapelet方法

本小节先形式化地给出多变量时间序列分类问题、shapelet的相关定义以及模拟退火算法的相关内容,再详细介绍本文所提的方法Multi-shapelet。

2.1　相关定义

2.1.1　多变量时间序列分类问题和shapelet的相关定义

一条多变量时间序列是一组序列数据,它常常被认为是在相等间隔的时间段内,依照给定的采样率、使用多个传感器对某个潜在过程进行观察的结果[18]。本文将一个多变量时间序列数据集定义为$T=\{(T_i, c_i): i=1, 2, \cdots, n\}$,其中$n$为$T$中多变量时间序列样本的个数,$c_i \in C$为样本$T_i$的类标,$C$为一个有限的类标集合。$T_i=\{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_L\} \in \mathbf{R}^{m \times L}$为$T$中的一个多变量时间序列样本,其中$m$为多变量时间序列样本的变量数,$L$为多变量时间序列样本的长度,$\boldsymbol{t}_p=\{t_{1,p}, t_{2,p}, \cdots, t_{m,p}\}', p \in \{1, 2, \cdots, L\}$为$T_i$中在时间点$p$的一个长度为$m$的向量。多变量时间序列分类任务就是对于一个给定的标签未知的多变量时间序列样本T_i,通过一定的方法预测其类标c_i。令$\boldsymbol{s}_b=\{t_{b,p}, t_{b,p+1}, \cdots, t_{b,p+l-1}\}, b \in \{1, 2, \cdots, m\}$代表样本$T_i$的一个属于变量$b$的长度为$l$的子序列。需要说明的是,本文提到的多变量时间序列的子序列均为从某个变量中提取的单变量序列。对于给定两个分别属于变量b_1和b_2的长度为l的子序列s_{b_1}和s_{b_2},s_{b_1}和s_{b_2}之间的欧式距离定义为

$$\mathrm{dist}(s_{b_1}, s_{b_2})=\sqrt{\frac{1}{l}\sum_{k=1}^{l}\left\| t_{b_1,k}-t_{b_2,k}\right\|^2} \tag{1}$$

对于一个长度为L的多变量时间序列样本T_i和长度为$l(l \leqslant L)$的子序列s_b,使用长度为l的滑动窗口在T_i上提取所有长度为l的子序列$\{s_{b,1}, s_{b,2}, \cdots, s_{b,L-l+1}\}, b=\{1, 2, \cdots, m\}$,则$\boldsymbol{s}_b$和$T_i$的距离定义为$\boldsymbol{s}_b$和$T_i$的所有长度为$l$的子序列的最小距离,即

$$D(\boldsymbol{s}_b, T_i)=\min_{\substack{\forall j \in \{1, 2, \cdots, m\},\\ k \in \{1, 2, \cdots, L-l+1\}}} \mathrm{dist}(\boldsymbol{s}_b, s_{j,k}) \tag{2}$$

式中$D(\cdot)$代表计算不同长度序列之间欧式距离的函数。

shapelet转换技术最早由Hills等[19]提出用于单变量时间序列分类问题,对于一个给定的单变量时间序列样本$\boldsymbol{X}=\{x_1, x_2, \cdots, x_m\}$和包含$k$条shapelet集合$S=\{s_1, s_2, \cdots, s_k\}, 1 \leqslant |s_i| \leqslant m$。shapelet转换技术可以将$\boldsymbol{X}$转化为一个包含$k$个特征值的特征向量$V=\{v_1, v_2, \cdots, v_k\}$,其中$v_i$代表$s_i$到$X$的距离,$v_i$可定义为

$$v_i=D(s_i, X) \tag{3}$$

本文用到的是文献[7]中提出的多变量shapelet转换技术。在多变量shapelet转换技术中,对于给

定的多变量时间序列样本 $T_i=\{\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_L\}\in\mathbf{R}^{m\times L}$ 和 shapelet 集合 $S=\{s_{b_1,1},s_{b_2,2},\cdots,s_{b_k,k}\}$，其中 $1\leqslant|s_{b_i,i}|\leqslant L$，$s_{b_i,i}$ 代表该 shapelet 从属于变量 b_i 的序列中提取出来的子序列。多变量 shapelet 转换技术根据候选 shapelet 集合 S 将 T_i 转换成长度为 k 的特征向量 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_k\}$。与单变量 shapelet 转换不同的是，在多变量 shapelet 转换中，特征值 v_i 是通过计算 $s_{b_i,i}$ 与多变量时间序列 T_i 中属于变量 b_i 的时间序列的距离得到的[20]。图 3 展示了一个多变量 shapelet 转换的实例，对于给定的变量数为 6 的多变量时间序列样本 T_j 和 3 条 shapelet S_1、S_2、S_3，通过多变量 shapelet 转换技术，将样本 T_i 转换了成长度为 3 的特征向量。

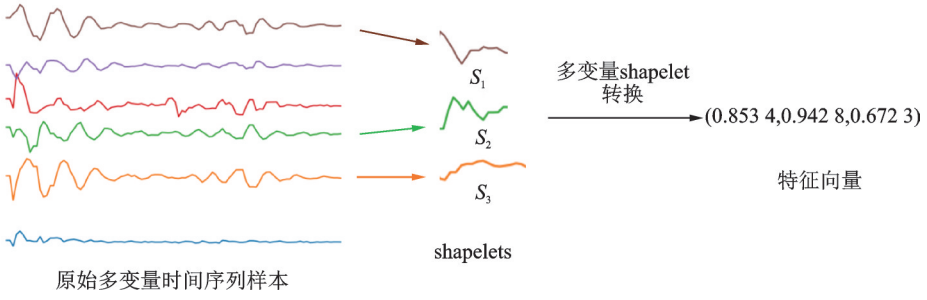


图 3 多变量 shapelet 转换示例

Fig.3 Multivariable shapelet transformation example

2.1.2 模拟退火算法

模拟退火算法是 20 世纪 80 年代早期发展起来的一种随机性组合优化方法[21]，其出发点来自热力学与统计物理学中的固体退火过程与一般的组合优化问题之间的相似性。

模拟退火在进行优化时先确定初始温度，随机选择一个初始状态并考察该状态的目标函数值；对当前状态附加一小扰动，并计算新状态的目标函数值；以概率 p 接受较好点，以概率 $1-p$ 接受较差点作为当前点，直到系统冷却。由于模拟退火算法以某种概率接受较差点，因而具有跳出局部最优解的能力。模拟退火方法在初始温度足够高、温度下降足够慢的条件下，能够以概率意义上收敛到全局最优值[22]，因而该方法在工程中得到了广泛的应用[23-24]。

2.2 分类模型设计

Multi-shapelet 的整体框架如图 4 所示，该方法分为切片、编码、shapelet 剪枝、shapelet 转换、shapelet 选择和使用标准分类器分类 6 个步骤。

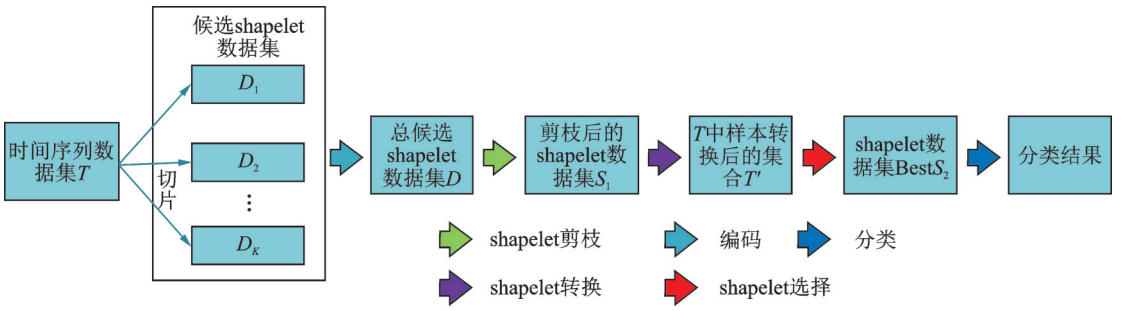


图 4 基于无监督表示学习和 shapelet 的多变量时间序列分类方法的整体框架

Fig.4 Overall framework of multivariate time series classification method based on unsupervised representation learning and shapelet

2.2.1　切片

在同一个数据集中,每个时间序列样本通常会包含很多非常相似的子序列,而能够决定一条时间序列类别的,往往是这条时间序列中的某条子序列[25],shapelet就是这样一种能够在很大程度上决定某条时间序列所属类别的子序列,其长度往往只占一条完整时间序列的小部分。因此为了从时间序列数据集中提取候选shapelet,使用滑动窗口对时间序列进行切片,将原始时间序列样本切割成一条条给定长度的小片段。本文根据文献[6,17]中对切片长度的设置将滑动窗口的大小按比例设置为[0.2L, 0.4L, 0.6L],其中L为原始时间序列的长度。这样,将得到3个具有不同长度的候选shapelet集合$D_i, i=1,2,3$。

2.2.2　编码

基于shapelet的分类方法通常需要从大量候选shapelet的集合中筛选分类效果好的shapelet用于后续的分类,这种筛选往往需要在不同长度和变量的候选shapelet之间进行比较。然而,来自不同变量长度不同的候选shapelet之间通常很难比较,因此,本文提出使用融合扩张卷积神经网络和图神经网络(Dilated convolution neural network and graph neural network, DC-GNN)混合模型作为编码器,将长度不同的候选shapelet映射到维度一致的选择空间,消除不同变量和不同长度的候选shapelet之间难以比较的问题,这样每个不同长度的候选shapelet都有一个长度统一的嵌入。

DC-GNN模型的详细结构如图5所示,该模型由两个模块和一个全局平均池化层构成,第一个是扩张因果卷积网络模块,该模块由多个残差块构成,其中每个残差块的结构图5中也同样给出,每个残差块中包含多个因果卷积网络层,通过堆叠在一起的残差块,扩张因果卷积神经网络可以有效提取来自原始时间序列不同尺度的信息;第二个是图神经网络模块,该模块包括两个图卷积层,其作用在于加快在原空间中正/负样本点对应的嵌入的聚集速度,减少模型的训练轮数和时间。全局平均池化层除了起到连接上述两个模块的作用外,还是DC-GNN模型能够处理长度不同的序列并将其映射成长度统一的嵌入的关键。

为了训练DC-GNN模型,本文在基于簇的三元组损失函数的基础上提出了一个新的损失函数使其可以达到两个目的:目的一为约束映射后正/负样本点对应的嵌入之间的距离相较于原空间正/负样本

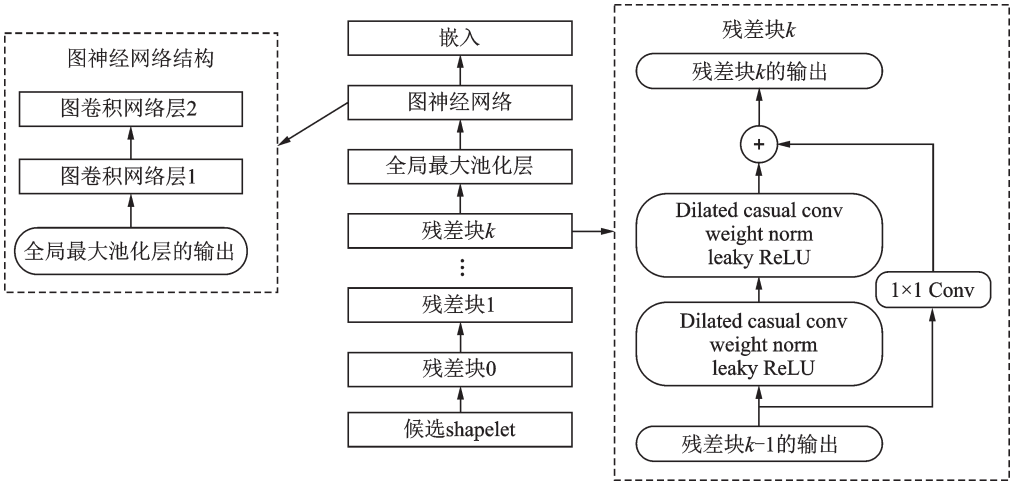


图5　DC-GNN模型的详细结构图

Fig.5　Detailed structure diagram of DC-GNN model

点之间的距离更紧密以及正样本点对应的嵌入同负样本点对应的嵌入的距离更远;目的二为约束正/负样本点对应的嵌入中每对嵌入之间的距离与原空间中对应的每对正/负样本点之间距离的变化率尽可能的相似。其中,实现目的一的原理是为了使原空间中属于同类的候选shapelet在新空间中的距离更紧密,这有助于后续基于K-means的shapelet剪枝过程。实现目的二原理为:在使用编码器学到候选shapelet对应的嵌入后,接着使用基于K-means聚类的剪枝方法,根据shapelet对应的嵌入将其聚成K类,从中选出K个离类中心最近的嵌入,这里嵌入与候选shapelet是一一对应的。因此,只有编码后嵌入的拓扑分布和原始候选shapelet分布的变化率尽可能的相似,才能保证在新空间中具有代表性的嵌入所对应的shapelet在原空间的分布中同样具有代表性。如果这点无法保证,基于K-means算法选出的具有代表性的嵌入对应的shapelet在候选shapelet集合中无法保证具有代表性,这会影响后续的分类精度。

本文所提出的损失函数如式(4)所示。由式(4)可以看出损失函数Loss由两个部分组成,其中α和β为超参数,$L1$是Li等[6]提出的基于锚节点、正样本簇和负样本簇的三元损失函数,其中正样本簇和负样本簇中的样本点可通过对候选shapelet进行聚类来确定,通过选择正/负样本点中离类中心最近的一些点作为模型训练使用的正/负样本簇被证明对模型的收敛是有效的[6],而锚节点则是从正样本簇中随机选的一个点。

$$\mathrm{Loss}=\alpha*L1+\beta*\lg(L2+1) \tag{4}$$

$L1$如式(5)所示,其由分别由D_{AP}、D_{NP}以及D_{intra}构成,D_{AP}和D_{NP}分别如式(6)和式(7)所示,其中K^+和K^-分别代表正样本簇和负样本簇中样本点的数量,(x,x^+,x^-)分别代表锚节点、正样本点和负样本点,f指代编码器的映射。D_{AP}和D_{NP}分别代表锚节点到正样本簇和负样本簇的距离。如式(8)所示,D_{intra}代表正样本簇中距离最大的样本点嵌入的距离值与负样本簇内距离最大的两个样本点嵌入的距离值之和。由D_{AP}、D_{NP}、D_{intra}构成的$L1$已被文献[6]证明能有效地达到目的一,$L1$的值越小,越能保证目的一的实现。

$$L1=\lg\left(\frac{D_{AP}+\mu}{D_{AN}}\right)+\lambda D_{\mathrm{intra}} \tag{5}$$

$$D_{AP}=\frac{1}{K^+}\sum_{i=1}^{K^+}\mathrm{dist}(f(x),f(x_i^+)) \tag{6}$$

$$D_{NP}=\frac{1}{K^-}\sum_{i=1}^{K^-}\mathrm{dist}(f(x),f(x_i^-)) \tag{7}$$

$$D_{\mathrm{intra}}=\max_{1\leqslant i<j\leqslant K^+}\mathrm{dist}(f(x_i^+),f(x_j^+))+\max_{1\leqslant i<j\leqslant K^-}\mathrm{dist}(f(x_i^-),f(x_j^-)) \tag{8}$$

$L2$的提出主要是为了完成目的二,$L2$的公式如式(9)所示,其同样由两项构成,第1项为正样本簇映射后对应的嵌入之间的距离值与原空间相应样本点的距离值比值的方差,第2项为负样本簇映射后样本点对应的嵌入之间的距离值与映射前样本点距离值比值的方差,这两项之和构成了$L2$。$L2$的值越小,映射前后拓扑的变化越接近等比例的变化,也就越能保证目的二的实现。$L1$与$L2$的加权和构成的Loss越小,越能保证目的一和目的二的实现,这样也就能使映射前后同类样本之间的拓扑变化为等比例的缩小。

$$L2=\frac{1}{(K^+-1)^2}\sum_{1\leqslant i<j\leqslant K^+}(D_{ij}-\overline{D^+})^2+\frac{1}{(K^--1)^2}\sum_{1\leqslant i<j\leqslant K^-}(D_{ij}-\overline{D^-})^2 \tag{9}$$

$$D_{ij}=\frac{\mathrm{dist}(f(x_i),f(x_j))}{\mathrm{dist}(x_i,x_j)} \tag{10}$$

$$\overline{D^{\pm}}=\frac{1}{(K^{\pm}-1)^2}\sum_{1\leqslant i<j\leqslant K^{\pm}}D_{ij} \tag{11}$$

2.2.3　shapelet剪枝

在通过编码器学习到候选shapelet的嵌入后,还需要通过这些嵌入构成的集合筛选分类效果最好的shapelet集合。这样做主要出于以下两个方面的考虑:(1)如果考虑所有的候选shapelet进行分类,这种做法不光在计算上是不可接受的,还很难得到分类的最优解;(2)处于同一个数据集中的时间序列样本包含着大量相似的候选shapelet,而时间序列的类间变化通常又只取决于少部分的候选shapelet。因此对候选shapelet集合进行一定的剪枝和筛选操作十分必要,本文分别使用K-means聚类方法和模拟退火算法实现对候选shapelet的剪枝和筛选。

使用K-means聚类算法对候选shapelet进行剪枝操作的原理在于,时间序列数据集通常包含很多分布在不同实例上的相似模式,而来自同一类的序列通常遵循相似的模式,故相似的模式会在同一类的时间序列中重复[25],因此可以通过使用K-means聚类算法过滤掉相似的候选shapelet。首先将候选shapelet数据集在新空间对应的嵌入聚成N类,其中N为人为指定的超参数;其次将这N类中离类中心最近的嵌入所对应的候选shapelet筛选出来,将其余候选的shapelet丢弃;最后得到了含有N个候选shapelet的集合,这里将其记为$S_1=\{s_{b_1,1},s_{b_2,2},\cdots,s_{b_N,N}\}$,其中$s_{b_i,i}$代表该shapelet是从属于变量$b_i$的时间序列中提取出来的。

由于前一步的剪枝操作只能完成从候选shapelet集中选出有代表性的N个候选shapelet集S_1,而不是分类能力最强的N个shapelet组合,因此为了找到分类能力最强的shapelet组合,在得到包含N个候选shapelet的集合S_1后,还需进行下一步的筛选。需要另外说明的是,在利用shapelet对时间序列进行分类时,每一个shapelet相当于是时间序列的一个特征,而每一个时间序列样本到该shapelet的距离就是该样本在该特征上的特征值,那么从包含N个候选shapelet的集合S_1中筛选出最终的K个分类能力强的shapelet集合就转换成了一个特征选择问题,在这里K是人为指定的超参数,因此可以选择使用一些有效的特征选择方法进行筛选,而基于模拟退火算法的特征选择方法[26]就是这样一种有效的特征选择方法。

2.2.4　shapelet转换

在使用特征选择算法之前,为了节省特征选择算法的运行时间,本文使用多变量shapelet转换技术根据S_1将多变量时间序列数据集转化为了长度为N的特征向量,避免后续的重复转换操作。图3给出了多变量shapelet转换的示意图,如图3所示,经过多变量shapelet转换后,一个多变量时间序列样本就转换成了长度等于shapelet数量的特征向量。

2.2.5　shapelet选择和分类

在使用多变量shapelet转换技术将多变量时间序列样本都转换成长度为N的特征向量后,接下来使用基于模拟退火算法的特征选择方法从S_1中选出分类效果较强的K个shapelet集合BestS_2,本文使用的基于模拟退火算法的特征选择方法的流程如图6所示。需要说明的是,流程图中所描述的BestS_2、$S_{2'}$以及S_2不是直接包含K个shapelet的集合,而是包含K个shapelet在S_1中的索引的集合,为了方便说明,直接使用BestS_2、S_2'以及S_2指代包含K个shapelet的集合。

如图6所示,基于模拟退火算法的特征选择方法首先通过随机从S_1中随机选择K个shapelet初始化候选shapelet集合S_2并给出初始温度T_0,接着计算基于S_2在训练集上的分类精度CA,为了防止过拟合,在这里使用了k折交叉验证法,验证S_2在训练集上的分类精度,实验中此处k的值被设置为3。由于

支持向量机在分类问题上良好的性能，这里分类精度的计算是通过支持向量机来完成的，其中支持向量机的相关参数设置与文献[6]中的参数设置保持一致。接着在S_2的基础上进行更新，将更新后的shapelet集合记为S_2'，并计算S_2'在训练集上的分类精度NA，若NA小于CA，则根据式(12)生成一个概率值P。

$$P = e^{(NA-CA)/CT} \tag{12}$$

式中：CT代表当前温度，P代表向差方向移动的概率。可以看出，温度越高，向差方向移动的概率越大，否则令$P=0$。之后生成一个0～1之间的随机数r，并判断是否满足$NA>CA$或$r<P$，若满足则接受S_2'，并令$S_2=S_2'$和$CA=NA$，接着判断NA是否大于BA，若满足，则保存当前分类精度最好的shapelet集合S_2，并令Best$S_2=S_2$和$BA=CA$，最后判断当前温度是否小于终止温度T'，若满足则返回分类精度最好shapelet集合BestS_2，否则令$CT=\omega*CT$，其中ω是指定的温度衰减系数，然后回到Step 3，接着搜索分类精度更好的shapelet集合，直到得到分类能力最强的包含K个shapelet的集合BestS_2后，使用多变量shapelet转换技术将测试集中的样本转换为长度为K的特征向量，接着使用支持向量机进行分类，得到shapelet集合BestS_2在测试集上的分类精度。

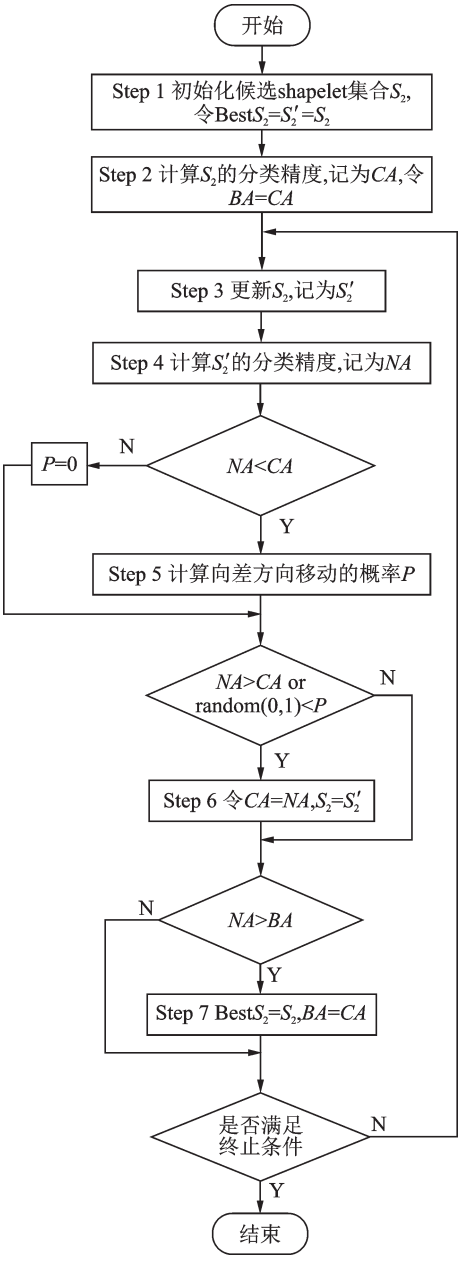


图6 基于模拟退火算法的特征选择方法

Fig.6 Feature selection method based on simulated annealing algorithm

3 实 验

3.1 实验运行环境

使用python实现本文所提方法Multi-shapelet及6种基准方法，所有实验均在Xeon E5-2630 V4@2.20 GHz和4张2080Ti配置的Ubuntu系统的机器上运行。

3.2 数据集和参数

在著名的UEA[27]中18个多变量时间序列数据集上测试了所提出的方法。数据集的相关统计信息如表1所示，其中UEA数据集是多变量时间序列分类问题中的一个基准数据集，包括来自人类活动识别、人类发音识别、心电图及脑磁图信号识别等多个领域的数据集，代表性较强，故常被用于测试新的方法。在Multi-shapelet及对比方法的实验中，UEA中所有的多变量时间序列数据集均按照8∶2的比例被划分为了训练集和测试集。在使用UEA的数据集之前，针对部分数据集中存在的缺失值，使用了0值填充，之后各个数据集中的样本进行了相应的标准化操作，标准化过程为

$$\widehat{t_{b,i}} = \frac{t_{b,i} - \overline{t_b}}{\sigma_b} \tag{13}$$

式中:$t_{b,i}$为时序变量b的第i个值,$\overline{t_b}$和σ_b分别为时序变量b的均值和标准差,$\widehat{t_{b,i}}$为$t_{b,i}$标准化之后的值。

Multi-shapelet方法中的DC-GNN模型的扩展因果卷积模块的超参数同文献[6]设置的一样,学习率、批处理大小、通道数、卷积核的大小、K-means聚类的类别数N以及网络的深度分别被设置为0.001、10、40、3、100和10。最终筛选出的shapelet的数量K被设置为50,式(5)中μ和λ的值分别被设置为0.2和1,α和β的值分别被设置为0.9和0.1,初始温度T_0、终止温度T'和温度衰减系数ω分别被设置为50、1和0.95,模型的训练轮数被设置为25,相比于文献[6]中同类方法的400轮训练次数,本文所提的shapelet表示学习模型明显更加高效。

表1　UEA中18个多变量时间序列数据集的相关统计信息

Table 1　Relevant statistics for 18 multivariate time series datasets in UEA

数据集	大小	样本维度	序列长度	样本类别
ArticularyWordRecognition	575	9	144	25
AtrialFibrillation	30	2	640	3
BasicMotions	80	6	100	4
Cricket	180	6	1 197	12
DuckDuckGeese	100	1 345	270	5
Epilepsy	275	3	206	4
ERing	300	4	65	6
Handwriting	1 000	3	152	26
Heartbeat	409	61	405	2
JapaneseVowels	640	12	29	9
Libras	360	2	45	15
LSST	4 925	6	36	14
NATOPS	360	24	51	6
RacketSports	303	6	30	4
SelfRegulationSCP1	561	6	896	2
SelfRegulationSCP2	380	7	1 152	2
StandWalkJump	27	4	2 500	3
UWaveGestureLibrary	440	3	315	8

3.3　对比方法

为了验证Multi-shapelet的分类效果,将Multi-shapelet同当前分类效果较好的6种方法进行了比较,下面简要介绍一下这些方法。

(1)EDI和DTWI[14]:基于ED和DTW距离度量方法的多变量时间序列分类方法。

(2)MLSTM-FCNs[9]:采用LSTM层和堆叠的CNN层来提取时间序列的特征,接着根据提取的特征用softmax层进行分类。

(3)ResCNN[8]:该方法在CNN的基础上结合残差网络提取时间序列特征,而后使用softmax层进行分类。

(4)TST[10]:该方法基于transformer提取序列特征,而后使用该特征进行分类。

(5)ShapeNet[6]:该方法利用三元损失函数训练编码器学习候选shapelet的新表示,而后基于这个新表示在shapelet剪枝、shapelet发现的基础上使用支持向量机进行分类。

3.4　实验结果

Multi-shapelet同其他方法在UEA18个多变量时间序列数据集上的分类精度对比结果如表2所示。由表2可以发现,Multi-shapelet在14个数据集中表现最好,且在12个数据集的精度超过了ShapeNet方法。此外,与其他方法相比,在大多数数据集中,Multi-shapelet同样具有显著的优势,这表明Multi-shapelet具有一定的普适性。综上所述,不难看出本文方法Multi-shapelet的总体准确度是所有比较方法中最好的,这或许是因为这些数据集中的确存在有效的shapelet,且本文方法的确能找到分类精度较高的shapelet集合。

接下来对Multi-shapelet与其他6种基线方法在UEA数据集上的分类精度进行了Wilcoxon符号秩检验,以检测Multi-shapelet与其他基线方法在18个数据集中的结果是否存在显著的差异,表2给出了Wilcoxon符号秩检验的结果。Wilcoxon符号秩检验的结果表明,Multi-shapelet与其他方法在$p\leqslant0.05$

表2 本文方法同基准方法在UEA中的18个多变量时间序列数据集的分类精度

Table 2 Classification accuracy of the proposed method and benchmark methods on 18 multivariate time series datasets in UEA

数据集	EDI	DTWI	MLSTM-FCNs	Res-CNN	TST	Shape-Net	Our method
ArticularyWordRecognition	0.965	0.987	0.967	0.974	0.974	0.987	**0.991**
AtrialFibrillation	0.167	0.200	0.267	0.333	0.167	0.167	**0.500**
BasicMotions	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Cricket	0.917	**1.000**	0.986	**1.000**	**1.000**	0.986	**1.000**
DuckDuckGeese	0.450	0.400	0.650	0.650	0.550	0.650	**0.700**
Epilepsy	0.564	0.927	0.971	0.964	0.964	**0.982**	**0.982**
ERing	0.950	0.915	0.852	0.933	**0.967**	0.950	**0.967**
Handwriting	0.550	0.610	0.272	**0.710**	0.540	0.610	0.580
Heartbeat	0.683	0.678	0.741	0.756	0.585	0.756	**0.781**
JapaneseVowels	0.969	0.930	**0.992**	**0.992**	**0.992**	**0.992**	0.984
Libras	0.847	0.883	0.839	**0.903**	0.847	0.875	**0.903**
LSST	0.507	0.556	**0.644**	0.384	0.503	0.587	0.575
NATOPS	0.875	0.839	0.933	**0.944**	0.917	**0.944**	**0.944**
RacketSports	0.869	0.809	0.875	0.836	0.803	0.875	**0.918**
SelfRegulationSCP1	0.841	0.775	0.782	**0.884**	0.792	0.867	**0.884**
SelfRegulationSCP2	0.447	0.539	0.506	0.474	0.685	0.789	**0.868**
StandWalkJump	0.333	0.200	0.400	0.167	0.400	0.400	**0.833**
UWaveGestureLibrary	0.898	0.903	0.753	0.932	**0.943**	0.920	0.932
Total best acc	1.000	2.000	3.000	7.000	5.000	4.000	14.000
Ours 1-to-1-Wins	17.000	15.000	15.000	10.000	13.000	12.000	—
Ours 1-to-1-Draws	1.000	2.000	1.000	6.000	3.000	3.000	—
Ours 1-to-1-Losses	0.000	1.000	2.000	2.000	2.000	3.000	—
RankMean	5.220	4.670	4.060	2.940	3.830	2.440	1.560
Wilcoxon Test p-values	0.000	0.001	0.002	0.001	0.021	0.012	—

的情况下均存在显著差异。最终实验结果表明,Multi-shapelet的精度显著优于其他方法。

为了验证本文所提的损失函数在训练模型时的收敛性能,本文给出了在AtrialFibrillation、StandWalkJump、BasicMotions以及UWaveGesture-Library这4个数据集上的损失函数随着训练次数增加的变化情况,具体如图7所示。从图7可以看出,本文所提损失函数在约束属于同类样本对应的嵌入之间的相对位置形成的拓扑与原空间中样本点之间的相对位置形成的拓扑之间的关系更接近于一种等比例的缩小这一目标时,收敛性能同样可以得到保证,且在训练轮次较少时,损失值就已经比较低了,这表明本文所提混合模型DC-GNN具有较高的学习效率。

3.5 消融实验

为了研究本文所提的新损失函数对最终分类精度的影响,分别使用三元组损失函数和本文提出的损失函数来训练DC-GNN模型,图8给出了这两种方式训练的模型在4个数据集中的分类效果,其中这

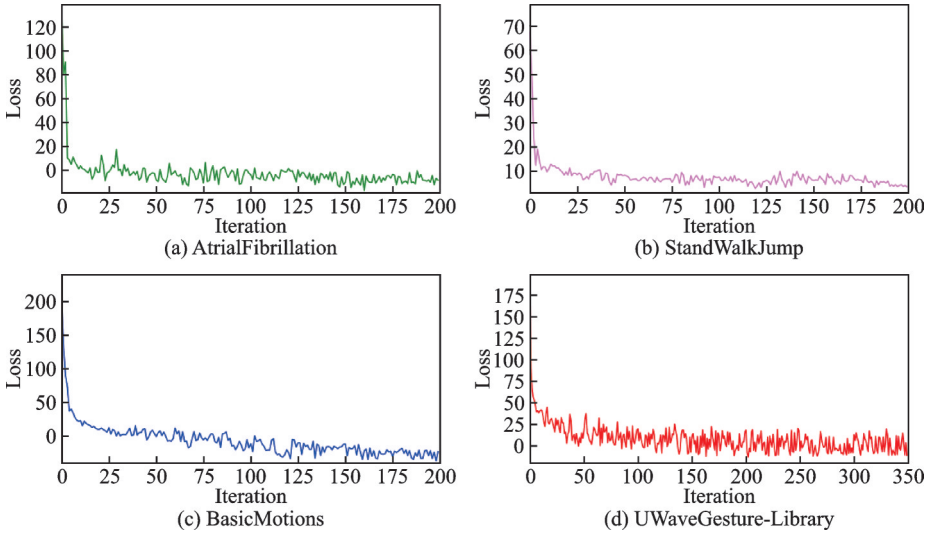


图7　本文所提损失函数在4种不同数据集上的收敛情况

Fig.7　Convergence of the proposed loss function on four different data sets

4个数据集分别是ArticularyWordRecognition、NATOPS、Libras和Heartbeat。由图8可以看出,本文所提损失函数的表现优于三元损失函数的表现,这说明本文所提损失函数相比于三元损失函数更有利于约束模型学到好的嵌入从而提高最终分类精度。

为了研究本文所提出的特征选择方法对最终分类精度的影响,在其他条件相同的情况下,在shapelet选择阶段分别使用基于模拟退火算法的筛选方法和文献[6]中提出的固定选择函数对shapelet进行筛选。图9给出了这两种选择方法在ArticularyWordRecognition、NATOPS、Libras和Heartbeat这4个数据集中的分类效果。由图9可以看出,本文使用的基于模拟退火算法的特征选择方法的表现超过了使用固定选择函数的表现,且对比图8可以发现,特征选择方法的改进对于最终分类精度的提升帮助更大。

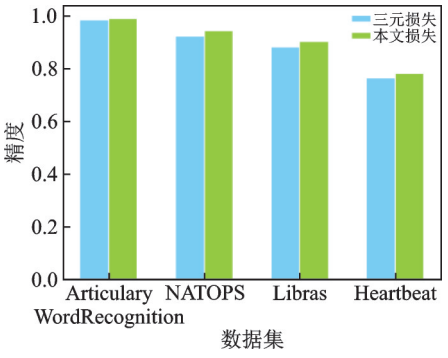


图8　三元组损失函数和本文所提新损失函数的分类效果对比

Fig.8　Comparison of classification effect between triple loss function and the proposed new loss function

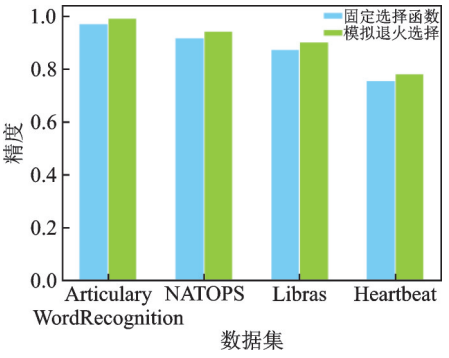


图9　固定选择函数同基于模拟退火的特征选择方法的分类效果对比

Fig.9　Comparison of classification effect between the fixed selection function and the proposed feature selection method based on simulated annealing

3.6 可解释性实验

相比于其他方法，基于shapelet的分类方法天然具备一定的可解释性，然而这些可解释性通常需要一定的专业知识，为了更好说明本文所提方法Multi-shapelet的可解释性，在所需专业知识较少的BasicMotions数据集上进行了实验。

BasicMotions中收集了人类在进行4种不同运动时身上所佩戴的3D加速计和3D陀螺仪产生的共计6个维度的时间序列样本，其中这4种运动为站立、走动、跑动和打羽毛球，分别用S、W、R和B来指代。

图10给出了基于shapelet在BasicMotions数据集进行多变量shapelet转换的可视化案例，其中s_1和s_2是从样本中提取的两个shapelet，分别来自样本的第2维和第6维，接着进行shapelet转换，将转换后的特征在坐标轴上进行可视化。图10左侧给出了不同类别的样本实例示意图。图10右侧中坐标点的x值和y值分别是s_1和s_2转换而来的特征值，根据坐标轴中的可视化结果可以发现，不同种类的运动存在着较为明显的聚集效益，且站立S和打羽毛球B距离较近，这和人们的常识相符。因为在打羽毛球的过程中，存在着较长时间的站立，而跑动R与走动W距离较近，与站立S距离最远，这同样符合人们关于这3项运动的常识。根据以上分析可以看出，本文所提方法Multi-shapelet具备一定程度的可解释性。

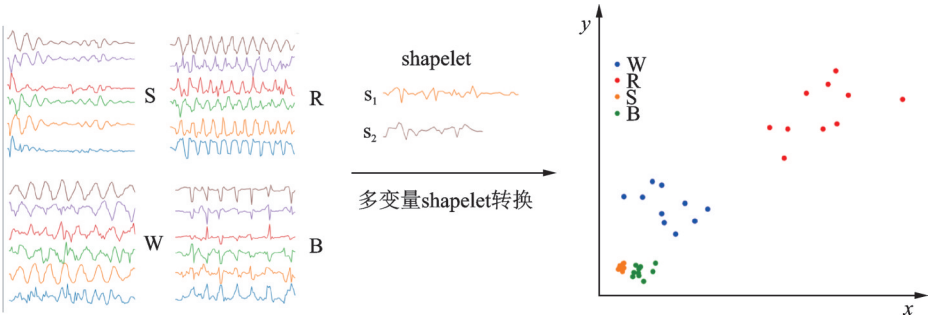


图10 在BasicMotions数据集上的多变量shapelet转换案例

Fig.10 Multivariable shapelet transformation case on BasicMotions dataset

4 结束语

本文提出了一种基于无监督表示学习和shapelet的多变量时间序列分类方法Multi-shapelet。首先通过混合模型DC-GNN作为编码器将不同长度的候选shapelet嵌入统一的shapelet选择空间，以进行shapelet之间的比较；其次，提出了一种损失函数以无监督学习的方式训练该编码器，使得DC-GNN对样本编码后，属于同类样本对应嵌入之间相对位置形成的拓扑与原空间中样本点之间相对位置形成的拓扑之间的关系更接近于一种等比例的缩小，这对后续基于相似性的剪枝过程十分重要；接着，使用K-means聚类和模拟退火算法进行shapelet剪枝和选择操作，选出分类能力较强的shapelet集合。在UEA的18个多变量时间序列数据集上的实验结果表明，Multi-shapelet在可解释性和精度上，相比于其他方法具有显著优势。未来将通过探索更高效的编码器，使学到的表示在shapelet发现算法中发挥更好的作用，从而进一步提高模型的分类精度和效率。

参考文献：

[1] YE L, KEOGH E. Time series shapelets: A new primitive for data mining[C]//Proceedings of the 15th ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2009: 947-956.

[2] KARLSSON I, PAPAPETROU P, BOSTRÖM H. Forests of randomized shapelet trees[C]//Proceedings of International Symposium on Statistical Learning and Data Sciences. [S.l.]: Springer, 2015: 126-136.

[3] MUEEN A, KEOGH E, YOUNG N. Logical-shapelets: An expressive primitive for time series classification[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2011: 1154-1162.

[4] YAN Xinming, MENG Fanrong, YAN Qiuyan. Shapelet classification method based on trend feature representation[J]. Journal of Computer Applications, 2017, 37(8): 2343.

[5] MA Qianli, ZHUANG Wanqing, LI Sen, et al. Adversarial dynamic shapelet networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2020, 34(4): 5069-5076.

[6] LI Guozhong, CHOI B, XU J, et al. Shapenet: A shapelet-neural network approach for multivariate time series classification [C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021, 35(9): 8375-8383.

[7] BOSTROM A, BAGNALL A. A shapelet transform for multivariate time series classification[EB/OL]. [2017-12-18]. https://doi.org/10.48550/arXiv.1712.06428.

[8] ZOU X, WANG Z, LI Q, et al. Integration of residual network and convolutional neural network along with various activation functions and global pooling for time series classification[J]. Neurocomputing, 2019, 367: 39-45.

[9] KARIM F, MAJUMDAR S, DARABI H, et al. Multivariate LSTM-FCNs for time series classification[J]. Neural Networks, 2019, 116: 237-245.

[10] ZERVEAS G, JAYARAMAN S, PATEL D, et al. A transformer-based framework for multivariate time series representation learning[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. [S.l.]: ACM, 2021: 2114-2124.

[11] HUANG Xin, KHETAN A, CVITKOVIC M, et al. Tabtransformer: Tabular data modeling using contextual embeddings [EB/OL]. [2017-12-11]. https://arxiv.org/abs/2012.06678v1.

[12] KEOGH E, WEI L, XI X, et al. Supporting exact indexing of arbitrarily rotated shapes and periodic time series under Euclidean and warping distance measures[J]. The VLDB Journal, 2009, 18(3): 611-630.

[13] MEI J Y, LIU M, WANG Y F, et al. Learning a Mahalanobis distance-based dynamic time warping measure for multivariate time series classification[J]. IEEE Transactions on Cybernetics, 2015, 46(6): 1363-1374.

[14] SHOKOOHI-YEKTA M, WANG J, KEOGH E. On the non-trivial generalization of dynamic time warping to the multi-dimensional case[C]//Proceedings of the 2015 SIAM international conference on data mining. [S.l]: Society for Industrial and Applied Mathematics, 2015: 289-297.

[15] YE L, KEOGH E. Time series shapelets: A novel technique that allows accurate, interpretable and fast classification[J]. Data Mining & Knowledge Discovery, 2011, 22(1/2): 149-182.

[16] SCHÄFER P, LESER U. Multivariate time series classification with WEASEL+MUSE[EB/OL]. [2018-08-17]. https://doi.org/10.48550/arXiv.1711.11343.

[17] FRANCESCHI J Y, DIEULEVEUT A, JAGGI M. Unsupervised scalable representation learning for multivariate time series [J]. Advances in Neural Information Processing Systems, 2019. DOI:10.48550/arXiv.1901.10738.

[18] 赵慧赟,潘志松.基于shapelets学习的多元时间序列分类[J].计算机科学,2018,45(5): 180-184, 219.
ZHAO Huiyun, PAN Zhisong. Multivariate time series classification based on shapelets learning[J]. Computer Science,2018,45(5): 180-184, 219.

[19] HILLS J, LINES J, BARANAUSKAS E, et al. Classification of time series by shapelet transformation[J]. Data Mining and Knowledge Discovery, 2014, 28(4): 851-881.

[20] 闫汶和,李桂玲.基于shapelet的时间序列分类研究[J].计算机科学,2019, 46(1): 29-35.
YAN Wenhe, LI Guiling. Research on time series classification based on shapelet[J]. Computer Science, 2019, 46(1): 29-35.

[21] KIRKPATRICK S, GELATT JR C D, VECCHI M P. Optimization by simulated annealing[J]. Science, 1983, 220(4598): 671-680.

[22] VAN LAARHOVEN P J M, AARTS E H L. Simulated annealing[M]. Netherlands: Springer, 1987: 7-15.

[23] 李元香,项正龙,夏界宁.模拟退火算法的动力系统模型及收敛性分析[J].计算机学报,2019,42(6): 1161-1173.
LI Yuanxiang, XIANG Zhenglong, XIA Jiening. Dynamical system models and convergence analysis for simulated annealing algorithm[J]. Journal of Computers, 2019,42(6): 1161-1173.

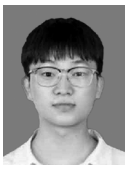
[24] 杨玮,李然,张堃.基于变邻域模拟退火算法的多自动导引车任务分配优化[J].计算机应用,2021,41(10): 3056-3062.
YANG Wei, LI Ran, ZHANG Kun. Task allocation optimization for automated guided vehicles based on variable neighborhood simulated annealing algorithm[J]. Computer Application,2021,41(10): 3056-3062.

[25] GRABOCKA J, WISTUBA M, SCHMIDT-THIEME L. Fast classification of univariate and multivariate time series through shapelet discovery[J]. Knowledge and Information Systems, 2016, 49(2): 429-454.

[26] ABDEL-BASSET M, DING W, EL-SHAHAT D. A hybrid Harris Hawks optimization algorithm with simulated annealing for feature selection[J].Artificial Intelligence Review, 2021, 54(1): 593-637.

[27] BAGNALL A, DAU H A, LINES J, et al. The UEA multivariate time series classification archive, 2018[EB/OL]. [2018-10-31]. https://doi.org/10.48550/arXiv.1811.00075.

作者简介:



詹熙(1999-),男,硕士研究生,研究方向:机器学习和时间序列分析,E-mail:1013230453@qq.com。

黎维(1995-),男,博士研究生,研究方向:机器学习和时空序列预测。

潘志松(1973-),**通信作者**,男,教授,博士生导师,研究方向:机器学习和计算机视觉,E-mail:panzhisong@aeu.edu.cn。

(编辑:王静)