计算机科学
COMPUTER SCIENCE

基于Transformer特征融合的时间序列分类网络

段梦梦，金城

引用本文

段梦梦，金城. 基于Transformer特征融合的时间序列分类网络[J]. 计算机科学，2023，50(12)：97-103.

DUAN Mengmeng, JIN Cheng. Transformer Feature Fusion Network for Time Series Classification[J]. Computer Science, 2023, 50(12): 97-103.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100112

基于 Transformer 特征融合的时间序列分类网络

段梦梦[1]　金　城[2]
1 复旦大学软件学院　上海 200438
2 复旦大学计算机科学技术学院　上海 200438
(mmduan20@fudan.edu.cn)

摘　要　在时间序列分类任务中,模型集成方法通过训练多个基础模型并利用一定的规则来聚合基础模型的输出,从而得到比单一基础模型更准确的结果。目前模型集成方法主要关注基础模型的选择以及如何提高基础模型的差异性和多样性,忽视了对聚合规则的探索。针对这一问题,提出了基于 Transformer 特征融合的时间序列分类网络(Transformer Feature Fusion Network,TFFN)。该网络包含二重 Transformer 编解码器(Dual Transformer Encoder Decoder,Dual TED)和基于 Transformer 的具有样本分布感知特性的分类模块(Transformer Encoder Head,TEH)两个核心组件。Dual TED 利用 Transformer 的注意力模块对基础特征进行提取和融合,得到具有更强辨别性的融合特征。具有样本分布感知特性的分类模块根据融合特征对时间序列进行更准确的分类,从而弥补现有集成模型方法忽视特征融合、集成规则过于简单的不足。实验结果表明,TFFN 在多个主流时间序列分类数据集上取得了最好的成绩。

关键词:时间序列分类;模型集成;Transformer;特征融合;深度学习
中图法分类号　TP391

Transformer Feature Fusion Network for Time Series Classification

DUAN Mengmeng[1] and JIN Cheng[2]
1 School of Software,Fudan University,Shanghai 200438,China
2 School of Computer Science,Fudan University,Shanghai 200438,China

Abstract　Model ensemble methods train multiple basic models and use a certain rule to aggregate the output of the basic models for time series classification. However,they mainly focus on two aspects. The first one is which model is chose as the basic model. And the Second one is how to increase the difference and the diversity of the basic models. They all ignore the exploration of aggregation rules. Aiming at this problem,Transformer feature fusion network for time series classification(TFFN) is proposed. TFFN have two key components,dual Transformer encoder decoder(Dual TED) and Transformer encoder head(TEH). Dual TED leverage attention module to fuse the basic feature into more discriminative fusion features. Transformer encoder head,a sample-distribution-aware classifier,is adopted to classify time series more accurately. Experiments show that TFFN achieves state-of-the-art results on multiple mainstream time series classification datasets.

Keywords　Time series classification,Model ensemble,Transformer,Feature fusion,Deep learning

1　引言

时间序列是按数值产生的时间顺序进行排列的数值序列,如股票数据和居民用电数据等。多年以来,时间序列分类一直是数据挖掘领域的基础问题,相关研究成果在金融、气象、工程等多个领域得到了越来越广泛的应用。随着传感器等数据采集技术的发展,时间序列的类型和内容也在不断丰富,数据随时间变化的规律也更加复杂多样,时间序列分类也因此面临新的挑战。

时间序列分类的难点主要在于不同序列或不同变量在时间尺度上的差异较大,导致时域特征难以提取。目前单一分类模型[1-3]通常只针对某些类型的时间序列有较好的分类效果,在泛化能力上仍然存在着较大不足。

一些研究者采用模型集成的思路,将多个基础模型集成在一起,从而综合不同模型的优点以获得更好的分类效果。早期模型集成方法[4-5]将多个基础模型得到的分类概率向量进行投票或加权平均,来得到最终的分类结果。近期的一些方法[6-7]注意到,由辨别性特征生成分类概率向量的过程丢失了大量信息,转而采用直接对特征进行聚合的方法来获得更好的分类效果。然而,这些方法的特征聚合过程比较简单,当

到稿日期:2022-11-11　返修日期:2023-03-30
基金项目:上海市科技创新行动计划(22dz1204900)
This work was supported by the Shanghai Municipal Science and Technology Commission(22dz1204900).
通信作者:金城(jc@fudan.edu.cn)

基础模型差异较大时，特征难以被有效地聚合，因此分类能力提升有限。

近年来，计算机视觉、自然语言处理等领域常使用基于Transformer的特征融合技术[8-9]。得益于注意力机制，来自不同模态的、差异较大的特征仍然能被有效地融合。鉴于此，本文结合Transformer和模型集成方法的优点，提出了基于Transformer特征融合的时间序列分类网络(Transformer Feature Fusion Network，TFFN)。该网络利用Transformer的注意力机制对基础特征进行提取和融合，再利用具有样本分布感知特性的分类模块，根据融合特征对时间序列进行分类，从而弥补现有模型集成方法忽视特征融合、集成规则过于简单的不足。本文工作的主要贡献如下：

1)提出了一种基于Transformer特征融合的时间序列分类网络TFFN，其核心模块是一种二重编解码器结构(Dual Transformer Encoder Decoder，Dual TED)。该结构能够有效地对基础模型所提取的特征进行双向融合，形成更具辨别性的融合特征，提升特征的表达能力。

2)使用具有样本特征分布感知的分类模块(Transformer Encoder Head，TEH)，依据样本特征在全体特征中的相对位置(而非其在特征空间中的绝对位置)对样本进行分类，从而提高分类准确率。

3)在两个主流的时间序列分类数据集上进行了充分的实验。结果表明，本文方法具有最优秀的分类能力。

2 相关工作

2.1 单一分类模型

近年来，单一分类模型的发展主要围绕如何解决不同时间序列在时间尺度上差异较大、可辨别性特征提取方式难统一的问题。Schäfer等[10]针对变量间时间尺度差异大的难点，提出了WEASEL+MUSE方法来提取多个变量之间的整体性特征。Fawaz等[11]针对时间尺度难确定的问题，将Inception模块引入到时间序列分类任务中，提出了InceptionTime网络。Tang等[12]提出了一种全尺度卷积网络OS-CNN，通过对不同大小的卷积操作进行加权求和，来适应不同数据集下时间尺度的差异。

另一些研究则关注提升特征的表达能力。Zhang等[13]提出了一种基于注意力的原型网络TapNet，从时间序列中提取隐含特征。针对训练标签有限的问题，TapNet利用未标注数据与标注原型样本之间的距离来半监督地训练特征表示。Dempster等[14]针对特征单一的问题，提出了一种利用大量随机大小和随机权重的卷积核来提取时域信息的方法ROCKET。之后，Dempster等[15]又提出了一种改良版的ROCKET，称为MiniRocket方法。该方法采用一组小且固定的卷积核取代了原方法中的随机卷积核，降低了模型的计算复杂度。Zerveas等[16]则将Transformer引入时间序列分类领域，采用几何分布掩码，在训练过程中通过随机遮挡来进行数据增强，从而提升分类网络的特征表达能力。由于不同时间序列之间的差异较大，分类依据也不尽相同，因此单一模型难以同时对多个不同类型的时间序列进行有效的分类。

2.2 模型集成

区别于单一分类模型，模型集成的方法通过训练多个基础模型并根据一定的规则来聚合基础模型的输出，从而得到比单一基础模型更准确的分类结果。模型集成方法在多个领域的分类问题上都有应用，例如Wortsman等[17]在图像分类领域提出了Model Soups模型。该模型采用了贪心算法的思想，将已知较优的若干个模型按照准确率排序，然后依次尝试将其加入到集成的模型中，然后将其在ImageNet上的准确率作为权重进行加权求和。

在时间序列分类问题上，也有研究者使用了模型集成方法。Lines等[18]将多种类型的传统方法集成在一起，构建了一个基于层次化投票的集成方法HIVE-COTE，集成了包含辨别子序列、全序列距离度量等多种方法。Shifaz等[19]在HIVE-COTE的基础上，提出了TS-CHIEF方法。该方法采用了弹性相似性度量、频谱特征提取等技术，在保持分类精度的同时缩短了模型训练和推理的时间。Middlehurst等[20]在HIVE-COTE的基础上提出了HIVE-COTE 2.0。该方法替换了原有的分类器，采用了时态字典集成和多样表示规范区间森林，以此降低分类耗时，并将ROCKET的集成分类器作为该方法的一个集成组件，以此提高分类准确度。但是，HIVE-COTE2.0与其原始版本一样，都存在只能处理定长时间序列的缺点。

从整体上看，基于模型集成的方法主要关注基础模型的选择，而忽视对聚合规则的探索。使用简单加权平均或者投票来决定分类结果的策略[21-22]往往使得集成模型的性能反而不如基础模型(特别是基础模型分类准确率差异较大时)。一些研究者试图构建网络融合基础特征来得到分类结果。Ishaq等[23]针对时序依赖关系提取困难的问题，提出了一种利用全连接网络进行特征融合的模型集成方法。Xia等[24]针对模型对时间窗口参数设定敏感的问题，提出了一种包含多种时间窗口的集成方法，该方法利用一个双向LSTM方法对各模型特征进行融合。然而，这些方法没有考虑到基础模型特征差异大、难以简单融合的问题，导致其准确率提升有限。

3 基于Transformer的特征融合网络

3.1 问题描述

时间序列是将采样结果按照时间排列而成的序列，根据采样需求的不同，采样结果包含一个或多个变量。一般地，时间序列可定义为：

$$\mathbb{S}=[\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_T],\boldsymbol{S}_i\in R^{N_s},i=1,2,\cdots,T \tag{1}$$

其中，T表示时间序列$\mathbb{S}$的总长度，N_s表示时间序列的变量数，$\boldsymbol{S}_i$表示第i个采样数据。

时间序列分类问题，即依据某个规则f对时间序列进行分类$c=f(\mathbb{S})$，得到分类标签$c\in C$。其中，$C=\{C_i\}_{i=1}^{n}$表示所有分类标签的集合，n表示所有可能的类别总数。

3.2 整体结构

基于Transformer融合特征的时间序列分类网络(TFFN)从结构上看大致可以分为特征提取、特征融合和分类3个模块(见图1)。

特征提取模块先将基础模型输出的特征进行重投影，将它们映射到同一个特征空间中，再使用层归一化提高特征的一致性和稳定性，最后通过一个参数可学习的位置编码为其提供额外的位置信息。特征融合模块包含3个分支，每个

分支中均有一个二重 Transformer 编解码器（Dual Transformer Encoder- Decoder，Dual TED）。Dual TED 负责两个重投影特征之间的双向特征融合。3 个分支的特征通过全连接层变换到同一特征空间后被拼接在一起，得到融合后的特征。分类模块（Transformer Encoder Head，TEH）采用样本分布感知的分类器来进行分类，该分类器通过引入 Class Token 来学习全体样本的分布，并利用 Transformer Encoder 结合样本分布和融合特征，获取特征在样本分布中的相对位置，进而得到时间序列分类结果。

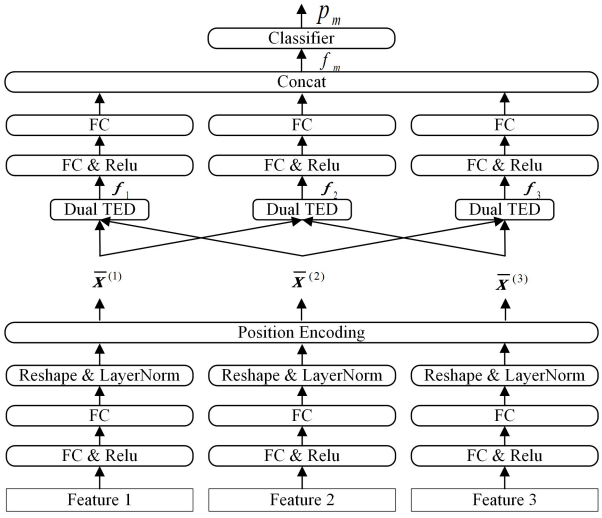

图 1　TFFN 的基本结构

Fig. 1　Structure of TFFN

3.3　特征提取模块

TFFN 首先使用 3 个既有时间序列分类模型来对输入时间序列提取初步特征，如果基础模型本身是一个分类模型，则取其最后一个 Softmax 层的输入作为其输出的基础特征。反之，如果基础模型只是一个特征提取器，则取该模型的输出作为基础特征。

对于给定的时间序列，不同的基础模型所提取的特征显然不同，特征之间的差异不仅仅体现在数值或者特征维度上，从本质上讲，这些特征属于不同的特征空间。对于不同空间下的特征，直接进行融合（如简单相加）显然缺乏理论依据，因此难以取得较好的效果。为此，TFFN 分别用两个全连接层对每个基础特征进行嵌入，即重投影到相同的特征空间（d_m 维）。

对于嵌入后的 d_m 维特征向量，TFFN 以每 d_t 维为一组，称为一个 Token，共计 N_t 个 Token，即把 d_m 维向量重新排列为 $N_t * d_t$ 的矩阵 $\boldsymbol{X}^{(i)}$。随后，TFFN 在 $\boldsymbol{X}^{(i)}$ 上添加了可学习位置编码 $\boldsymbol{P}^{(i)} \in R^{N_t * d_t}$，$\boldsymbol{P}^{(i)}$ 在模型训练之初被初始化为 0，位置编码过程可以表示为：

$$\bar{\boldsymbol{X}}^{(i)} = \boldsymbol{X}^{(i)} + \boldsymbol{P}^{(i)}, i \in \{1,2,3\} \tag{2}$$

其中，$\bar{\boldsymbol{X}}^{(i)}$ 即为第 i 个基础特征经过重投影和位置编码后的结果，而 3 个 $\bar{\boldsymbol{X}}^{(i)}$ 即是特征融合模块的输入。

3.4　特征融合模块

传统的模型集成方法采用简单拼接或是全连接网络来对特征进行融合，在基础特征差异较大的情况下融合效果欠佳。另一方面，计算机视觉领域中使用 Transformer 进行特征融合，取得了良好的效果。受这类方法的启发，本文设计了基于 Transformer 的特征融合模块，用于解决特征差异大导致融合困难的问题。

传统 Transformer 结构在特征融合上具有单向性的特点，即将一个特征融合进另一个特征中，也就是有一个特征作为主导。这种单向融合会导致非主导特征的部分信息在融合过程中丢失。而在 TFFN 中，3 个基础特征的关系是对等的，特征的融合也应该是相互的。为避免融合过程中的信息丢失，本文设计了一种二重编解码器模块（Dual TED），用于进行双向特征交流。

由于 Dual TED 模块一次只能融合两个特征，因此必须对 3 个基础特征 $\bar{\boldsymbol{X}}^{(i)}$ 进行两两融合，也就意味着 TFFN 中必须有 3 个融合分支，每个分支中均包含一个 Dual TED 模块。Dual TED 模块从结构上可以看作是由两组编解码器组成，其基本结构如图 2 所示。

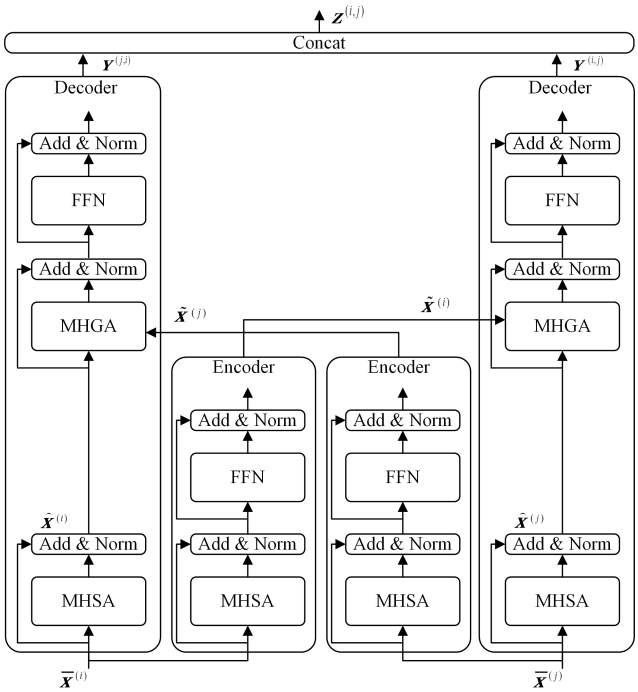

图 2　Dual TED 的结构

Fig. 2　Structure of Dual TED

对于输入的两个特征 $\bar{\boldsymbol{X}}^{(i)}, \bar{\boldsymbol{X}}^{(j)} \in R^{N_t * d_t}$，两组编解码器分别负责将特征 $\bar{\boldsymbol{X}}^{(i)}$ 融合到 $\bar{\boldsymbol{X}}^{(j)}$ 中和将特征 $\bar{\boldsymbol{X}}^{(j)}$ 融合到 $\bar{\boldsymbol{X}}^{(i)}$ 中。以前者为例，编解码器负责将特征 $\bar{\boldsymbol{X}}^{(i)}$ 融合到特征 $\bar{\boldsymbol{X}}^{(j)}$ 中，得到特征 $\boldsymbol{Y}^{(i,j)}$。特征 $\bar{\boldsymbol{X}}^{(i)}$ 即为融合方特征，特征 $\bar{\boldsymbol{X}}^{(j)}$ 即为被融合方特征，特征 $\boldsymbol{Y}^{(i,j)}$ 则为将特征 $\bar{\boldsymbol{X}}^{(i)}$ 融合到特征 $\bar{\boldsymbol{X}}^{(j)}$ 中的结果，即该编解码器的输出。Encoder 接受特征 $\hat{\boldsymbol{X}}^{(i)}$ 作为输入得到供 Decoder 提取的特征 $\tilde{\boldsymbol{X}}^{(i)}$，而 Decoder 接受特征 $\bar{\boldsymbol{X}}^{(j)}$ 和 Encoder 的输出 $\tilde{\boldsymbol{X}}^{(i)}$ 作为输入得到单向融合后的特征 $\boldsymbol{Y}^{(i,j)}$。

具体来说，Encoder 将特征 $\bar{\boldsymbol{X}}^{(i)}$ 通过一个多头自注意力模块（Multi-Head Self-Attention，MHSA），来提取 Token 之间的相关性，并根据相关性在 Token 层面进行交流和融合，从而得到 $\tilde{\boldsymbol{X}}'^{(i)}$。MHSA 提取的特征 $\tilde{\boldsymbol{X}}'^{(i)}$ 所处的特征空间与 $\bar{\boldsymbol{X}}^{(j)}$

不同，难以直接进行特征融合。因此，Encoder 将 $\widetilde{\boldsymbol{X}}'^{(i)}$ 送入一个由两个全连接层构成的前馈神经网络（Feed Forward Network，FFN）得到 $\widetilde{\boldsymbol{X}}^{(i)}$，使得 Encoder 的输出 $\widetilde{\boldsymbol{X}}^{(i)}$ 处于 $\overline{\boldsymbol{X}}^{(j)}$ 所在的特征空间中。本模块中所使用的 MHSA 和 FFN 均与标准 Transformer 中的结构一致。

Decoder 与 Encoder 的结构相似，其首先将特征 $\overline{\boldsymbol{X}}^{(j)}$ 通过一个 MHSA 得到 $\overset{\wedge}{\boldsymbol{X}}^{(j)}$。接下来，其中的多头引导注意力模块（Multi-Head Guided-Attention，MHGA）接受 $\widetilde{\boldsymbol{X}}^{(i)}$ 和 $\overset{\wedge}{\boldsymbol{X}}^{(j)}$ 作为输入，从 $\widetilde{\boldsymbol{X}}^{(i)}$ 中提取出与 $\overset{\wedge}{\boldsymbol{X}}^{(j)}$ 相关的特征，通过残差连接将相关特征融合到 $\overset{\wedge}{\boldsymbol{X}}^{(j)}$ 之中，得到 $\overline{\boldsymbol{Y}}^{(i,j)}$。其中，MHGA 的结构也与标准 Transformer 的结构一致。最后，Decoder 将 $\overline{\boldsymbol{Y}}^{(i,j)}$ 通过一个 FFN，来进行特征空间变换以及 Token 内部的特征交流，得到 $\boldsymbol{Y}^{(i,j)} \in R^{N_t * d_t}$。$\boldsymbol{Y}^{(i,j)}$ 即是一组编解码器的输出结果，也就是将特征 $\overline{\boldsymbol{X}}^{(i)}$ 融合到特征 $\overline{\boldsymbol{X}}^{(j)}$ 的结果。类似地，另一组编码器将特征 $\overline{\boldsymbol{X}}^{(j)}$ 融合到特征 $\overline{\boldsymbol{X}}^{(i)}$，得到融合结果 $\boldsymbol{Y}^{(j,i)}$。将两组编解码器的输出拼接在一起就得到了将 $\overline{\boldsymbol{X}}^{(i)}$ 和 $\overline{\boldsymbol{X}}^{(j)}$ 进行双向融合后的完整特征 $\boldsymbol{Z}^{(i,j)} \in R^{N_t * 2d_t}$。为方便后续表述，令 $\boldsymbol{f}_1$，$\boldsymbol{f}_2$，$\boldsymbol{f}_3$ 分别表示 3 个 Dual TED 输出的分支特征，即：

$$\boldsymbol{f}_1 = \boldsymbol{Z}^{(1,2)}, \boldsymbol{f}_2 = \boldsymbol{Z}^{(1,3)}, \boldsymbol{f}_3 = \boldsymbol{Z}^{(2,3)} \tag{3}$$

3 个分支特征在分别经过两层全连接层后被拼接到一起，从而得到了融合特征 $\boldsymbol{f}_m \in R^{3N_t * 2d_t}$。该特征将在分类模块中用于最终的时间序列分类。

3.5 分类模块

传统的分类方法通过样本特征在特征空间中所处的位置来对样本进行分类，这种分类方法基于一种假设——样本均匀分布在整个特征空间中。然而，这一假设并不总是成立。Dosovitskiy 等[25]通过研究发现，样本通常不会均匀地充斥整个特征空间，而总是聚集性地分布在特征空间中的某个区域，这导致仅利用绝对位置进行分类的准确率不高。基于这些发现，该文献提出了通过样本特征在全体特征所形成的特征空间中的相对位置来对样本进行分类的方法，即通过引入一组可学习的参数 $\boldsymbol{T}_c \in R^{1 * 2d_t}$，来学习全体样本的特征分布，该参数也称为 Class Token。受该项研究的启发，本文采用了一种基于 Transformer Encoder 的分类器（称为 TEH）。该分类器通过 Transformer Encoder 来结合全体样本的特征分布 $\boldsymbol{T}_c$ 和该样本的融合特征 $\boldsymbol{f}_m$，其结构如图 3 所示。

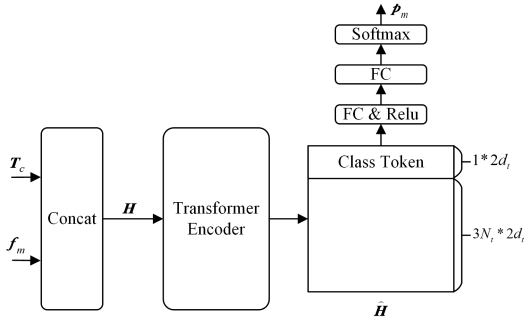

图 3 分类器 TEH 的结构

Fig. 3 Structure of TEH classifier

TEH 首先将 $\boldsymbol{T}_c$ 和 $\boldsymbol{f}_m$ 在 Token 维度进行拼接得到 $\boldsymbol{H} \in R^{(3N_t+1) * 2d_t}$，然后将其通过一个 Encoder 层进行融合得到输出 $\overset{\wedge}{\boldsymbol{H}} \in R^{(3N_t+1) * 2d_t}$。最后，将 $\overset{\wedge}{\boldsymbol{H}}$ 的前 $1 * 2d_t$ 维送入两个全连接层和一个 Softmax 层来转换成分类向量 $\boldsymbol{p}_m$。$\boldsymbol{p}_m$ 中数值最大的分量的索引即为分类结果 $c \in \{1, \cdots, N_c\}$。

3.6 损失函数

观察整个网络结构可以发现，关键性的特征共有 4 个，其中 3 个是来自不同 Dual TED 模块的分支特征，另 1 个则是最后的融合特征。因此，在设计损失函数时，一种直观的想法是希望能对这 4 个特征都进行约束，使得用其中每一个特征进行分类也能达到较高的准确率，从而推动网络同时学到 3 个 Dual TED 模块中特征信息的交互形式和最终的融合方法。

基于这种思路，TFFN 在训练阶段将 3 个分支特征 f_i 分别输出到 3 个分类器中，每个分类器的结构都是 2 个全连接层之后接 1 个 Softmax 层，则每个分支特征 f_i 经过分类器后都将输出一个分类概率向量 $\boldsymbol{p}_i$。如前文所述，在网络的分类阶段，融合特征 $\boldsymbol{f}_m$ 也会经过分类器输出分类概率向量 $\boldsymbol{p}_m$。这样，在训练过程中就会得到 4 个分类概率向量，将它们分别与样本标签求交叉熵即可得到它们各自的损失，即：

$$L_k = -\sum_{j=1}^{C} \boldsymbol{y}_j \log(\boldsymbol{p}_k^j), k \in \{m, 1, 2, 3\} \tag{4}$$

其中，C 表示分类数，$\boldsymbol{y}$ 表示样本真实标签，$\boldsymbol{y}_j$ 为其的第 j 个分量，$\boldsymbol{p}_k^j$ 表示分量概率向量 $\boldsymbol{p}_k$ 的第 j 个分量。

本文设计的总损失函数如下：

$$\mathcal{L} = L_m + \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 \tag{5}$$

其中，λ_1，λ_2，λ_3 都是权重值，用来平衡各个损失对总损失的贡献度。

4 实验

4.1 实验设置

本文实验的运行环境为 Ubuntu 20.04 64 位，1 个 Intel Xeon Silver 4210R CPU，128 GB 内存，512 GB 硬盘，1 块 Nvidia Titan RTX 3090 GPU，深度学习框架为 PyTorch1.8，Python 的版本为 3.8。

本文模型在训练时，嵌入层特征向量维度 d_m 设为 1 024，Transformer 编解码器中的节点数量 N_t 和节点维度 d_t 均设为 32，多头注意力机制中的 $h=8$。训练时采用 Adam 优化器，批次大小为 32，学习率为 0.001，共训练 100 个 epoch。

4.2 对比实验

为验证本文方法的有效性，在两个公开数据集 UCR128[26] 和 UEA[27] 上进行了对比实验。在性能评价方法上，与主流方法[26]保持一致，采用在所有子数据集上的平均排名作为评价指标。

4.2.1 UCR128 数据集

UCR128 数据集是由美国加利福尼亚大学河滨分校制作的单变量时间序列数据集，包括语音识别、健康监测和频谱分析在内的 128 个子数据集。UCR128 是 UCR85 数据集[28]的升级版本，相比后者，其挑战性有了较大的提高。UCR128 数据集的难度提升主要体现在 3 个方面：1）子数据集的数量

增加了 50%，覆盖范围也更广；2)在单个子数据集中，时间序列的长度不再保持一致，从而加大了模型训练的难度；3)采用了新的数据集划分策略，使得 UCR128 的测试集数据量大于训练集的数据量，从而对模型的泛化能力提出了更高的要求。

表 1 列出了本文方法同 7 个主流模型之间的性能对比。本文中，*Lose* 表示对比方法准确率更高的子数据集个数，*Win* 表示本文方法准确率更高的子数据集个数，*R* 表示该方法的平均排名。为充分论证本文方法的有效性，实验采用了较新的单一分类模型 OS-CNN，MiniRocket 和 InceptionTime 作为 TFFN 的基础模型进行集成，然后将 TFFN 在 UCR128 数据集上的分类结果与全部 7 个模型进行比较，完整的实验结果如表 1 所列。不难看出，TFFN 取得了最好的并且是远超其他方法的结果。

表 1 UCR128 数据集上的实验结果

Table 1 Comparison results on UCR128

方法名称	*Lose*	*Win*	*R*
TST[16]	30	**87**	5.77
PF[29]	34	**90**	5.59
OS-CNN[12]	47	**75**	4.59
InceptionTime[11]	39	**71**	4.36
MiniRocket[15]	30	**71**	4.32
TS-CHIEF[19]	53	**64**	3.92
ROCKET[14]	49	**70**	3.91
TFFN(本文方法)	—	—	**3.54**

此外，依据 Benavoli 等[30]的理论，采用 Wilcoxon-Holm post-hoc 分析计算各方法之间的显著性差异和平均排名，并以临界差异图(Critical Difference Diagram)的形式展示结果，如图 4 所示。

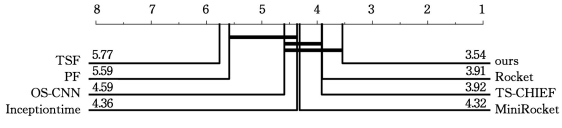

图 4 在 UCR128 上的临界差异图

Fig. 4 Critical difference diagram on UCR128

图 4 中，细折线上方数字表示相应方法的平均排名，横粗线表示其所连接的方法之间没有显著性差异。

4.2.2 UEA 数据集

UEA(University of East Anglia)是由英国东安格利亚大学和美国加利福尼亚大学河滨分校联合制作的多变量时间序列数据集，共有 30 个子数据集，包括人体运动、心电图、脑电图和声谱图分类等在内的多个领域。子数据集之间的差别较大，如分类数从 2 到 39 不等，多元变量的维度数最小的为 2，最大的为 963，时间序列的长度最短的只有 8，而最长的则为 17894。由于大部分对比方法仅提供了 26 个子集上的分类结果，因此本文也仅在这些子集上进行实验。

对比实验以 MiniRocket，TST 和 OS-CNN 为基础模型，采用本文方法进行集成得到融合模型，其与近期方法的对比实验结果如表 2 所列。不难看出，本文方法在单个子数据集上的表现全面占优，相比 2021 年以来提出的 3 种对比方法(即基础模型)，准确率领先的子数据集数量至少是对比方法的 2 倍，充分证实了本文方法的优越性。

表 2 UEA 数据集上的实验结果

Table 2 Comparison results on UEA

方法名称	Lose	Win	R
DTW-1NND(norm)[13]	3	**23**	6.38
DTW-1NND[13]	5	**21**	6.13
TapNet[13]	4	**21**	5.37
WEASEL+MUSE[10]	7	**17**	5.08
OS-CNN[12]	7	**18**	3.52
TST[16]	7	**15**	3.46
MiniRocket^{A} [15]	8	**17**	3.31
TFFN(本文方法)	—	—	**2.75**

本文方法同对比方法在 UEA 上的临界差异图如图 5 所示。

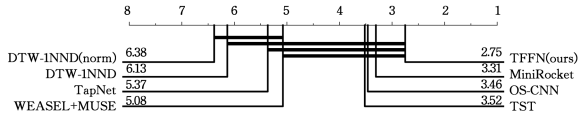

图 5 在 UEA 上的临界差异图

Fig. 5 Critical difference diagram on UEA

4.3 消融实验

本文方法的特点在于可学习位置编码、基于 Dual TED 的特征融合、TEH 分类器以及损失函数的设计。为验证上述技术点的必要性和有效性，在 UCR128 和 UEA 数据集上进行了关于聚合方法、损失函数、模型结构的消融实验。

4.3.1 聚合方法

采用多个 Dual TED 模块进行特征融合的设计思路在于，通过两两融合基础模型所产生的特征，使得基础特征之间能够充分地交换信息，从而在保留特征共有信息的同时渐近地提取出对分类有较强判别意义的新信息。为了论证这种聚合方法的有效性，采用了一种直观的替代方法来进行对比实验，即将基础模型产生的特征直接相加作为融合特征。考虑到基础模型所产生的特征所处的特征空间不同，其维度数也各不相同，因此在实验中保留了分类网络的特征映射阶段(使得特征融合在同一个特征空间中进行)，仅去掉了特征融合阶段，融合后的特征经过 Softmax 后输出分类概率向量。相应地，损失函数中去掉了 3 个分支的分类概率损失项，仅保留了融合特征的分类概率损失项。

另一种替代方案是将 3 个基础模型生成的特征分别通过 Softmax 层，获得其分类概率向量并求平均，将该结果作为最终的分类概率向量。两个替代方案的实验结果以及与本文方法的对比如表 3 所列。表 3 中，P 方法指对基础模型输出的分类概率向量求平均值后再计算分类标签的方法，F 方法则指对基础特征求和再进行分类的方法。

表 3 聚合方法的消融实验

Table 3 Ablation study of fusion methods

方法	UCR128			UEA		
	Lose	*Win*	*R*	*Lose*	*Win*	*R*
P	51	**53**	1.77	12	**14**	2.00
F	24	**85**	2.48	11	**15**	2.13
TFFN	—	—	**1.75**	—	—	**1.87**

可以看到，在融合方法上，采用本文方法的结果优于特征直接相加和对分类概率求平均的方法。这个结果表明，经过多分支融合获得的特征包含了更多更具判别性的信息，从而

使分类结果更加准确。值得注意的是，从分类结果上看，P 方法优于 F 方法，这也表明不加区分地对特征简单相加，会导致仅仅强化基础特征中的共性部分，没有充分利用基础特征中所包含的独有信息，从而实质上削弱了融合特征的表达能力，导致相加后特征的分类结果反而不如对分类概率求平均。

4.3.2 损失函数

本文损失函数设计思路包括两点：1）约束各分支的分类结果尽可能与真实值一致；2）约束融合特征的分类结果尽可能与真实值一致。直观上看，这两个约束似乎只要满足其中之一，另一个约束就自然会满足。但从其他领域的实践来看，上述观点并不正确，即满足其中一个约束条件并不必然导致另一个约束条件成立。为了验证本文所提损失函数在时间序列分类模型上的有效性，进行了两个消融实验，每个实验只约束上述两点之一。表 4 中的 LB 方法指损失函数中只保留分支特征的损失 L_1，L_2 和 L_3 的方法，LM 方法指损失函数中只保留 L_m 的方法。

表 4 损失函数的消融实验

Table 4 Ablation study of loss function

方法	UCR128			UEA		
	Lose	*Win*	*R*	*Lose*	*Win*	*R*
LB	42	**51**	1.90	10	**10**	1.80
LM	30	**70**	2.28	5	**18**	2.44
TFFN	—	—	**1.81**	—	—	**1.75**

不难发现，本文提出的分支分类损失结合融合分类损失的方法优于 LB 和 LM，这表明分支分类损失和融合分类损失在模型训练过程中起到的约束作用不尽相同，不能互相取代。此外，LB 方法的结果优于 LM 方法，说明对分支特征进行约束可以强化分支特征获取辨别性信息的能力，也说明了对特征进行两两融合思路的正确性。

4.3.3 模型结构

本文模型的结构主要有三大特点：1）相比传统 Transformer 编解码结构仅能将一个特征融合进另一个特征的做法，Dual TED 模块能够让两个特征相互从对方特征中提取出相关的部分，并进行双向特征融合；2）使用 Class Token 表示全体样本的特征分布，并使用 Transformer Encoder 将该特征分布和样本的融合特征相结合，得到该样本在全体样本中的相对位置；3）利用基于可学习参数的位置编码，为 Transformer 提供位置信息。本文进行了一系列的消融实验，以验证所提模型结构的有效性。各个替代方案的实验结果以及与本文方法的对比结果如表 5 所列。表 5 中所列的 TED 指采用传统 Transformer 编解码器替换 Dual TED 的方案，FCS 指采用两个全连接层和一个 Softmax 层替换 TEH 的方案，TF 指同时采用了 TED 和 FCS 的方案，PE 指移除位置编码的替代方案。可以看出，本文所提的 Dual TED 通过双向特征交流和融合的结构有效地提高了融合特征的辨别能力。相比传统分类方法，TEH 分类器在分类准确率上有较大提升。TFFN 中的位置编码也在一定程度上提升了模型的分类性能。无论是 TED 还是 FCS 都远强于 TF，这说明 Dual TED 和 TEH 从不同的方面强化了模型的分类能力，都是模型的重要组成部分。

表 5 模型结构的消融实验

Table 5 Ablation study of module structure

方法	UCR128			UEA		
	Lose	*Win*	*R*	*Lose*	*Win*	*R*
PE	32	**50**	2.75	10	**12**	2.63
TF	29	**76**	3.67	4	**19**	3.94
TED	40	**51**	2.68	8	**15**	2.94
FCS	32	**70**	3.30	11	**13**	2.96
TFFN	—	—	**2.59**	—	—	**2.52**

结束语 本文提出了一种基于 Transformer 融合特征的时序分类网络 TFFN，该网络利用所提的 Dual TED 结构能有效融合来自不同基础模型的基础特征，从而得到更具辨别性的融合特征。样本特征分布感知的分类模块 TEH 能根据融合特征在全体样本特征空间中的相对位置进行分类，提高了分类的准确性。对比实验结果表明，TFFN 在两个主流时间序列分类数据集上均取得了最优成绩。此外，本文还在聚合方法、损失函数和模型结构等方面进行了消融实验，充分证明了所提方法的有效性。

本文方法虽能有效地对基础特征进行融合，但 Dual TED 模块一次融合两个特征，使得模型较为复杂。当基础特征数量较多时，对特征进行两两融合的方式将使得计算量变得较大。因此，如何在较低计算量下一次性融合多个基础特征将是未来的研究方向。

参 考 文 献

[1] BAJWA M N, KHURRAM S, MUNIR M, et al. Confident classification using a hybrid between deterministic and probabilistic convolutional neural networks [J]. IEEE Access, 2020, 8: 115476-115485.

[2] LIU M H, ZENG A L, LAI Q X, et al. T-WaveNet: A Tree-Structured Wavelet Neural Network for Time Series Signal Analysis[C]// International Conference on Learning Representations(ICLR). 2021.

[3] XIAO Z, XU X, XING H, et al. RNTS: Robust Neural Temporal Search for Time Series Classification[C]// International Joint Conference on Neural Networks(IJCNN). 2021: 1-8.

[4] YAN W, LI G, WU Z, et al. Extracting diverse-shapelets for early classification on time series [J]. World Wide Web, 2020, 23(6): 3055-3081.

[5] LI H, JIA R, WAN X. Time series classification based on complex network[J]. Expert Systems with Applications, 2022, 194: 116502.

[6] YAGHOUBI V, CHENG L, VAN PAEPEGEM W, et al. An ensemble classifier for vibration-based quality monitoring[J]. Mechanical Systems and Signal Processing, 2022, 165: 108341.

[7] MELIN P, MONICA J C, SANCHEZ D, et al. Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico [J]. Healthcare, 2020, 8(2): 181.

[8] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]// Proceedings of the Conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019: 6558.

[9] YU Z，YU J，CUI Y，et al. Deep modular co-attention networks for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：6281-6290.

[10] SCHÄFER P，LESER U. Multivariate Time Series Classification with WEASEL+MUSE[EB/OL]. (2017-11) [2022-11]. https://arxiv. org/abs/1711. 11343.

[11] FAWAZ H I，LUCAS B，FORESTIER G，et al. InceptionTime：Finding AlexNet for time series classification[J]. Data Mining and Knowledge Discovery，2020，34(6)：1936-1962.

[12] TANG W S，LONG G D，LIU L，et al. Omni-Scale CNNs：A Simple and Effective Kernel Size Configuration for Time Series Classification[C]// International Conference on Learning Representations(ICLR). 2022.

[13] ZHANG X，GAO Y，LIN J，et al. TapNet：Multivariate Time Series Classification with Attentional Prototypical Network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：6845-6852.

[14] DEMPSTER A，PETITJEAN F，WEBB G I. ROCKET：exceptionally fast and accurate time series classification using random convolutional kernels[J]. Data Mining and Knowledge Discovery，2020，34(5)：1454-1495.

[15] DEMPSTER A，SCHMIDT D F，WEBB G I. MiniRocket：A Very Fast(Almost) Deterministic Transform for Time Series Classification[C]// The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD' 21). ACM，2021：248-257.

[16] ZERVEAS G，JAYARAMAN S，PATEL D，et al. A Transformer-based Framework for Multivariate Time Series Representation Learning[C]// The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD '21). ACM，2021：2114-2124.

[17] WORTSMAN M，ILHARCO G，GADRE S Y，et al. Model soups：averaging weights of multiple fine-tuned models improves accuracy without increasing inference time[C]// International Conference on Machine Learning. PMLR，2022：23965-23998.

[18] LINES J，TAYLOR S，BAGNALL A. Hive-cote：The hierarchical vote collective of transformation-based ensembles for time series classification[C]// 2016 IEEE 16th International Conference on Data Mining(ICDM). IEEE，2016：1041-1046.

[19] SHIFAZ A，PELLETIER C，PETITJEAN F，et al. TS-CHIEF：a scalable and accurate forest algorithm for time series classification[J]. Data Mining and Knowledge Discovery，2020，34(3)：742-775.

[20] MIDDLEHURST M，LARGE J，FLYNN M，et al. HIVE-COTE 2. 0：a new meta ensemble for time series classification[J]. Machine Learning，2021，110(11)：3211-3243.

[21] LI X，JIANG H，NIU M，et al. An enhanced selective ensemble deep learning method for rolling bearing fault diagnosis with beetle antennae search algorithm[J]. Mechanical Systems and Signal Processing，2020，142：106752.

[22] ZHOU Y，CHENG G，JIANG S，et al. Building an efficient intrusion detection system based on feature selection and ensemble classifier[J]. Computer Networks，2020，174：107247.

[23] ISHAQ M，KWON S. Short-term energy forecasting framework using an ensemble deep learning approach[J]. IEEE Access，2021，9：94262-94271.

[24] XIA T，SONG Y，ZHENG Y，et al. An ensemble framework based on convolutional bi-directional LSTM with multiple time windows for remaining useful life estimation[J]. Computers in Industry，2020，115：103182.

[25] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale[C]// International Conference on Learning Representations(ICLR). 2021.

[26] DAU H A，BAGNALL A，KAMGAR K，et al. The UCR Time Series Archive[J]. IEEE/CAA Journal of Automatica Sinica，2019，6(6)：6-18.

[27] BAGNALL A，DAU H A，LINES J，et al. The UEA multivariate time series classification archive，2018[A/OL]. [2018-10]. https://arxiv. org/abs/1811. 00075.

[28] CHEN Y P，EAMONN K，HU B，et al. The ucr time series classification archive[DS/OL]. [2015-07]. http://www. cs. ucr. edu/～eamonn/time_seires_data/.

[29] LUCAS B，SHIFAZ A，CHARLOTTE P，et al. Proximity forest：an effective and scalable distance-based classifier for time series[J]. Data Mining and Knowledge Discovery，2019，33(3)：607-635.

[30] BENAVOLI A，CORANI G，MANGILI F. Should we really use post-hoc tests based on mean-ranks? [J]. The Journal of Machine Learning Research，2016，17(1)：152-161.

DUAN Mengmeng，born in 1998，postgraduate. Her main research interest is time series classification and prediction.

JIN Cheng，born in 1978，Ph.D，professor，Ph.D supervisor，is a member of China Computer Federation. His main research interests include computer vision and multimedia information retrieval.

（责任编辑：喻藜）