计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(9):244−252 [doi: 10.15888/j.cnki.csa.009935] [CSTR: 32024.14.csa.009935]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于频域扰动的时间序列可解释性方法①

王景宇[1], 高德荃[2], 程佳俊[1,3]

[1](南京信息工程大学 软件学院, 南京 210044)
[2](国家电网有限公司信息通信分公司, 北京 100761)
[3](湖州师范学院 信息工程学院, 湖州 313000)
通信作者: 王景宇, E-mail: jingyuwang714@163.com

摘　要: 随着深度学习在时间序列分析中的广泛应用, 模型的预测性能得到了显著提升, 但其“黑箱”特性仍然限制了其在实际应用中的可信度与透明度. 目前, 虽然许多可解释性方法在一定程度上提供了对模型行为的洞察, 但它们在处理复杂时间序列数据时, 尤其是具有高频成分或长周期波动的任务中, 仍存在显著的局限性. 为解决这一问题, 本文提出了一种结合频域扰动和时间序列分段的时间序列可解释性方法. 通过短时傅里叶变换和动态时间规整算法, 本文首先对时间序列进行分段, 并通过生成对抗网络 (GAN) 生成频域扰动, 深入分析不同频率成分对模型预测结果的影响. 提出的方法能够在频域层面提供更精确的可解释性分析, 尤其在处理频率特征密集的时间序列任务中, 表现出明显优势. 实验结果表明, 所提方法在多个时间序列分类任务中均表现优越, 尤其在包含高频特征的任务中, 能够有效提高模型的可解释性并捕捉关键的频域特征. 与现有的基准方法相比, 我们的方法能够更精确地识别出影响预测结果的关键因素, 增强模型的透明度和可靠性.

关键词: 时间序列; 深度学习; 可解释性; 频域分析; 序列扰动分析

引用格式: 王景宇,高德荃,程佳俊.基于频域扰动的时间序列可解释性方法.计算机系统应用,2025,34(9):244–252. http://www.c-s-a.org.cn/1003-3254/9935.html

Time Series Interpretability Method Based on Frequency Domain Perturbation

WANG Jing-Yu[1], GAO De-Quan[2], CHENG Jia-Jun[1,3]

[1](School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](State Grid Information & Telecommunication Co. Ltd., Beijing 100761, China)
[3](School of Information Engineering, Huzhou University, Huzhou 313000, China)

Abstract: With the widespread application of deep learning in time series analysis, the predictive performance of models has been significantly improved. However, the “black-box” nature of these models still limits their trustworthiness and transparency in practical applications. Although many interpretability methods offer insights into model behavior, considerable limitations remain when handling complex time series data, particularly in tasks involving high-frequency components or long-period fluctuations. To address this challenge, this study proposes a time series interpretability method that integrates frequency-domain perturbation with time series segmentation. By employing short-time Fourier transform and dynamic time warping algorithms, the study first segments the time series, and then generates frequency-domain perturbations using generative adversarial network (GAN) to analyze the influence of different frequency components on model predictions. The proposed method enables more precise interpretability analysis at the frequency-domain level and demonstrates particular effectiveness in tasks characterized by dense frequency features. Experimental results show that the proposed method outperforms existing methods in multiple time series classification tasks,

① 基金项目: 国家电网信息通信分公司科技项目 (529939220001)
收稿时间: 2025-01-17; 修改时间: 2025-02-12; 采用时间: 2025-03-06; csa 在线出版时间: 2025-07-23
CNKI 网络首发时间: 2025-07-23

particularly in those involving high-frequency signals. It significantly improves model interpretability and captures key frequency-domain features. Compared to benchmark methods, the proposed approach more accurately identifies the critical factors influencing prediction outcomes, thus enhancing model transparency and reliability.

Key words: time series; deep learning; interpretability; frequency-domain analysis; sequence perturbation analysis

时间序列是指按照时间顺序排列的一组数据或观测值, 广泛存在于自然和社会现象中, 常见于金融[1]、气候[2]、医疗[3]、工业[4]和农业[5]等领域. 在时间序列数据中, 往往存在一定的时间依赖性, 表现为某一时刻的观测值与过去若干时刻的观测值间的相关性[6].

随着深度学习技术的发展, 基于神经网络的模型在时间序列预测领域崭露头角. 相较于传统的时间序列分析方法, 如自回归积分滑动平均 (ARIMA)[7]以及指数平滑法 (ETS)[8]等, 基于长短时记忆网络 (LSTM)[9]、门控循环单元 (GRU)[10]和 Transformer 架构[11]的模型凭借其卓越的特征学习能力与自动化的建模流程, 在复杂时间序列预测任务中取得了令人瞩目的突破. 这些模型能够充分挖掘大量历史数据中的潜在规律与模式, 进而实现更为精准的预测, 其预测精度相较于传统方法有了显著提升.

尽管深度学习模型在预测精度上表现优异, 其“黑箱”特性使得模型的内部机制和预测结果难以解释和理解. 因此, 如何提升这些复杂模型的可解释性, 成为当前时间序列分析研究中的一个重要课题. 通过增强模型的可解释性, 我们不仅可以更好地理解模型预测背后的原因和机制, 还能增加对预测结果的信任, 帮助领域专家做出更加合理的决策, 并进一步改进模型的性能和实际应用效果[12].

针对这一问题, 本文提出了一种基于频域分析的时间序列可解释性方法. 本文的贡献主要体现在以下几个方面.

(1) 通过短时傅里叶变换 (STFT), 将时间序列数据转化为频域表示, 提出了一种基于频域扰动的时间序列可解释性方法, 通过分析频域的视角来揭示时间序列模型中影响结果的关键因素.

(2) 设计了一种结合动态时间规整 (dynamic time warping, DTW) 和形状特征的算法对待解释时间序列进行分段, 并在分段后的序列中进行解释, 从而有效捕捉不同时间段对预测结果的影响.

(3) 在多个数据集上的实验结果表明, 所提方法能够有效提高模型的可解释性, 并取得了良好的表现.

1 相关研究

1.1 传统时间序列模型可解释性方法

近年来, 时间序列可解释性的研究逐渐成为一个重要的研究方向. 时间序列数据通常具有复杂的时序特性, 如何有效地解释和理解模型对这些数据的预测过程, 成为提升模型透明度和可靠性的关键. 部分可解释性方法依赖于梯度信息来衡量输入特征对模型输出的影响, 如积分梯度[13], 通过分析模型的梯度值来揭示哪些输入特征在预测过程中起到了重要作用.

基于注意力机制[14]的解释方法通过引入注意力层, 生成特征的重要性评分, 以识别时间序列中的关键时刻和特征. 这种方法能够较好地应对时间序列中不同时间点和特征之间的动态关系, 提供了比传统梯度方法更加灵活的可解释性.

此外, 扰动方法作为时间序列可解释性的常见方法, 得到了广泛研究. 扰动方法通常通过修改输入数据来评估特征对模型预测的影响, 尤其是在改变数据的某些部分后, 观察模型输出变化. 常见的扰动方法包括利用掩码进行扰动[15]、生成模型扰动[16]以及 LIME (local interpretable model-agnostic explanation)[17]等.

现有的时间序列可解释性方法在有效地揭示了输入特征对模型预测的影响中有不俗的表现, 能够较为灵活地应对时间序列中不同时间点和特征之间的动态关系. 然而, 在处理具有周期性或频率变化的时间序列时, 仅通过时域分析的可解释性效果并不理想.

1.2 基于频域分析的时间序列模型可解释性方法

频域分析是一种将时间序列从时间域转换到频率域的分析方法. 通过傅里叶变换或其他频域技术, 频域分析能够揭示数据中的频率成分和周期性模式. 通过对信号频率的分析, 可以帮助研究者识别出信号中的周期性波动、趋势变化等重要信息. 在时间序列分析中, 频域方法常用于分解复杂信号, 识别其中的周期模式和噪声成分[18].

与时域分析相比, 频域分析在处理周期性和长期趋势时更具优势. 时域分析通常依赖于直接观察时间序列数据的变化, 但在面对包含复杂周期性或频率变化的信号时, 时域分析往往难以有效捕捉到这些潜在的规律. 频域分析通过将时间序列转化为频率成分, 可以更清晰地揭示数据中的周期模式和频率相关的特征, 从而更好地识别数据中的周期性波动、趋势变化及噪声成分[19].

此外, 频域分析能够有效地从整体上分析信号的频率分布, 识别出对预测结果影响较大的频率成分. 与时域分析的逐时刻观察不同, 频域分析能够提供全局视角, 能够更加直观地揭示信号中的周期性波动.

频域分析在时间序列可解释性中的应用已经取得了一定进展, FreqRISE[20]基于频域掩码, 通过对频率成分的局部改变, 优化并生成可解释的特征. Vielhaben 等人[21]提出虚拟检查层 (virtual inspection layer), 利用傅里叶变换将时间序列数据映射到频域, 结合 LRP[22]方法, 为模型提供更具解释性的表示.

总的来说, 基于频域分析的时间序列可解释性方法能够将时间序列中的周期性波动和频率相关的特征从全局角度进行识别, 并有效过滤掉不相关的变化, 突出重要的频率成分, 揭示出数据中隐藏的周期性模式和长期趋势, 为模型提供更加直观和鲁棒的解释.

2 模型及方法

2.1 结合 DTW 和频域分析的时间序列分段方法

时间序列通常包含不同的阶段或状态, 这些阶段的特性可能在时间上有所不同, 如趋势、周期性或突发变化, 为了更好地捕捉时间序列中内在的模式和变化点, 设计了基于动态时间规整和频域分析的算法, 将原始序列划分为具有相似行为或统计特性的片段, 从而更好地捕捉时间序列中内在的模式和变化点. 对分段后的序列进行分析, 能够显著提升时间序列分析的效率与准确性, 将序列分解为同质区域, 减少了模型对全局噪声的敏感性, 并且分段能提供更直观的结果, 有助于解释复杂的时间序列数据.

方法具体内容如下, 为了实现时间序列分段, 将输入的时间序列$S=[s_1,s_2,\cdots,s_L]$, 划分为固定长度的重叠窗口, 每个窗口的长度为 W, 定义为:

$$S_i=[s_i,s_{i+1},\cdots,s_{i+w-1}],\quad i\in[1,L-w+1] \tag{1}$$

其中, S_i表示从序列S第i个位置开始、长度为 w 的子序列. 通过这样的窗口划分, 时间序列被分解为局部片段, 便于后续的相似性分析和模式识别.

接着, 通过计算每个窗口与其他窗口之间的动态时间规整 (DTW)[23]距离, 衡量不同时间窗口之间的非线性形状相似性. DTW 的目标是通过动态调整时间轴, 使得两个窗口之间的形状特性能够尽可能匹配, 其公式为:

$$DTW(S_i,S_j)=\min\sum_{k=1}^{K}\|S_i[k]-S_j[k']\|^2 \tag{2}$$

其中, k和k'分别表示序列 S_i, S_j中被对齐的时间点, K为对齐路径的长度. 通过计算所有窗口的 DTW 距离矩阵, 可以捕捉窗口之间的非线性关系.

计算过程中, 为每个窗口 S_i 找到其最近邻窗口$S_{nn(i)}$, 即与其在 DTW 意义下距离最小的窗口, 定义为:

$$S_{nn(i)}=\arg\min_j DTW\left(S_i,S_j\right),\quad j\neq i \tag{3}$$

通过这一计算, 每个窗口的最近邻信息被用来判断窗口间的相似性和一致性. 接下来, 检查相邻窗口之间的最近邻关系是否保持连续性. 如果相邻窗口S_i和S_{i+1}的最近邻关系中断, 即$S_{nn(i+1)}\neq S_{nn(i)}+1$, 则可以认为在位置$i$或$i+w-1$处的时间序列可能存在需要进一步分段解释的行为变化点. 在这种情况下, 将其标记为一个潜在的变点, 并记录下来作为后续分段的依据.

为了进一步提高变点检测的鲁棒性和准确性, 引入频域特性分析. 通过对候选变点的局部时间序列片段进行频域分析, 进一步对候选变点进行统计验证. 具体来说, 通过快速傅里叶变换 (FFT) 计算每个窗口的频谱特性$P(f)$, 并比较候选窗口与其后续窗口的频谱变化:

$$\rho_{\text{freq}}=\sum_f\left|P_{S_i}(f)-P_{S_{i+1}}(f)\right| \tag{4}$$

进一步将频域分析的结果与基于 DTW 计算的形状相似性结果ρ_{DTW}相结合综合考虑候选变点的显著性. 具体而言, 对于每个候选变点位置i, 同时利用ρ_{DTW}和ρ_{freq}来计算一个综合差异值ρ_{combined}, 衡量该位置作为变点的显著性. 综合差异值的计算公式为:

$$\rho_{\text{combined}}=\alpha\cdot\rho_{\text{DTW}}+\beta\cdot\rho_{\text{freq}} \tag{5}$$

其中, α和β是权重参数, 用于调节 DTW 差异和频域差异在综合评估中的重要性.

完成综合差异值的计算后, 对所有候选变点按照ρ_{combined}从大到小排序, 并选取综合差异值最大的T'个候选点作为最终的变点结果. 这些最终变点将时间序列划分为多个具有相似行为或统计特性的片段, 为后续的分析提供了更加清晰的结构化基础. 选取变点后的结果可以表示为: $\{c_1, c_2, \cdots, c_{T'}\}$, 其中$c_k$表示第$k$个变点的位置.

2.2 频域扰动生成

针对传统的扰动方法, 如随机噪声和掩码遮挡等, 其在对输入数据进行扰动时可能导致扰动后的数据出现分布外问题, 我们提出了一种基于生成对抗网络 (GAN) 的自适应扰动生成方法, 如图 1 所示, 首先对数据集使用短时傅里叶变换 (STFT) 从时域转换到频域, 接着使用生成对抗网络学习原始数据频域分布并生成分布内的扰动序列, 其中生成器损失由对抗损失和频域约束损失组成, 判别器损失函数由原始数据和生成数据分布差异以及梯度惩罚组成. 通过学习频域中的分布特性 (例如高频、低频及特定频率模式), 生成接近目标频域分布的数据. 本文采用 Wasserstein GAN (WGAN)[24] 的改进框架, 并且引入了梯度惩罚以满足 Lipschitz 连续性约束. 设计的生成对抗网络包括生成器G和判别器D. 生成器G的输入为随机噪声$z \sim N(0, I)$和条件向量T_{label}, 其中T_{label}, 表示目标频域分布的特性. 生成器通过映射噪声和条件向量生成时间域数据P_{gen}. 随后, 对生成数据P_{gen}应用 STFT, 得到其频域表示$F(P_{\text{gen}})$.

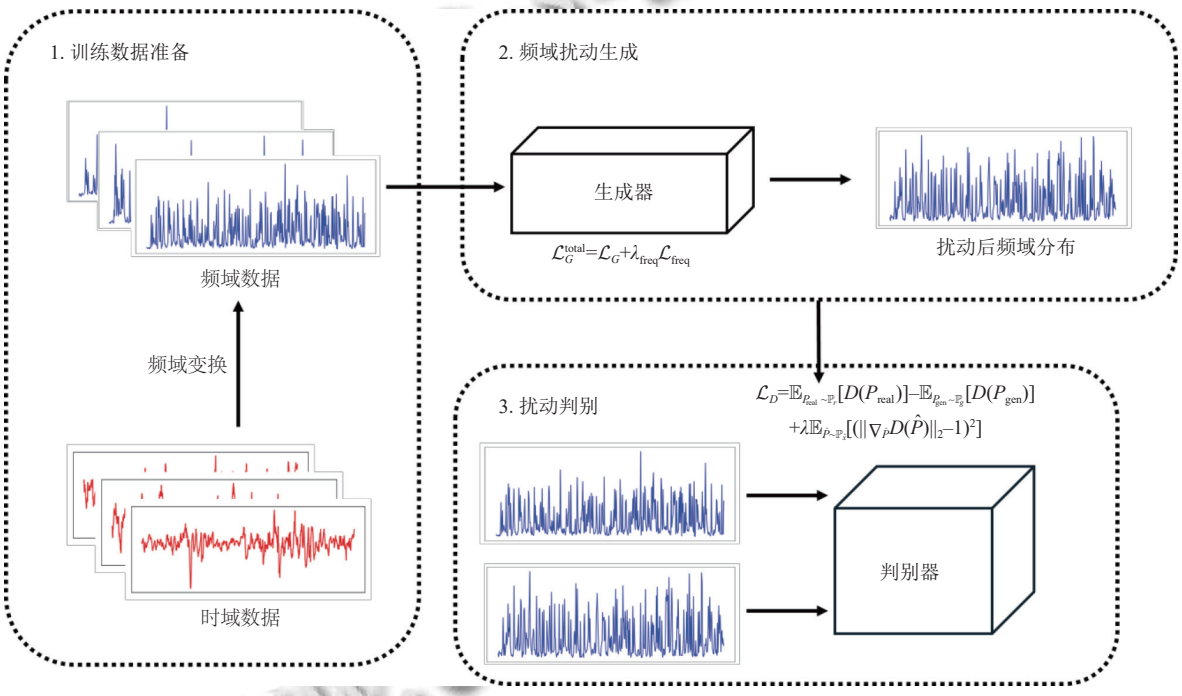

图 1 基于生成对抗网络的自适应频域扰动生成方法

为确保生成数据的频域特性接近目标分布T_{freq}, 判别器D的作用是区分输入数据是否为真实频域数据. 判别器接受真实频域数据P_{real}和生成频域数据P_{gen}作为输入, 判别器通过学习真实数据和生成数据的分布差异, 为生成器提供反馈, 指导其生成更接近真实数据分布的扰动数据. 为提升生成器和判别器的训练稳定性, 本文采用 WGAN 的改进框架, 通过优化 Wasserstein 距离替代传统 GAN 的对抗损失. 判别器直接输出一个标量值, 用于衡量数据的真实性, 而非通过概率分布进行二分类. 此外, 为满足 Lipschitz 连续性要求, 判别器引入了梯度惩罚机制, 通过约束判别器的梯度范数避免过拟合或不稳定训练. 判别器的损失函数定义为:

$$\mathcal{L}_D = \mathbb{E}_{P_{\text{real}}\sim\mathbb{P}_r}[D(P_{\text{real}})] - \mathbb{E}_{P_{\text{gen}}\sim\mathbb{P}_g}[D(P_{\text{gen}})] + \lambda\mathbb{E}_{\hat{P}\sim\mathbb{P}_{\hat{x}}}\left[(\| \nabla_{\hat{P}} D(\hat{P})\|_2 - 1)^2\right] \tag{6}$$

其中, $\mathbb{P}_r$和$\mathbb{P}_g$分别表示真实数据分布和生成数据分布, $\hat{P}$为从真实数据和生成数据之间的插值采样得到的样本, λ是梯度惩罚项的权重, 用于约束判别器的 Lipschitz 连续性.

生成器的优化目标包括两个部分. 首先, 生成器需要最小化对抗损失$\mathcal{L}_G$, 使生成数据尽可能被判别器判

定为真实数据:

$$\mathcal{L}_G = -\mathbb{E}_{P_{\text{gen}}\sim\mathbb{P}_g}[D(P_{\text{gen}})] \tag{7}$$

其次, 为了确保生成数据的频域特性$F(P_{\text{gen}})$与目标分布T_{freq}保持一致, 设计了频域约束损失函数:

$$\mathcal{L}_{\text{freq}} = \|F(P_{\text{gen}}) - T_{\text{freq}}\|_2^2 \tag{8}$$

其中, $\|\cdot\|_2^2$表示二范数, 用于衡量生成数据频谱与目标分布的均方误差, 通过引入频域约束, 我们期望生成的扰动序列能够保持与原始数据集相同的频域分布, 从而获得更为可靠的可解释性结果. 此外, 为进一步提升频域特性匹配的准确性, 本文引入了余弦相似性损失作为辅助约束:

$$\mathcal{L}_{\cos} = 1 - \frac{\langle F(P_{\text{gen}}), T_{\text{freq}}\rangle}{\|F(P_{\text{gen}})\| \cdot \|T_{\text{freq}}\|} \tag{9}$$

其中, $\langle\cdot,\cdot\rangle$表示向量的点积, $\|\cdot\|$表示向量模.

生成器的总损失函数是对抗损失和频域约束损失的加权和, 其定义为:

$$\mathcal{L}_G^{\text{total}} = \mathcal{L}_G + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}} \tag{10}$$

其中, λ_{freq}是权重参数, 用于平衡对抗损失与频域约束损失的影响.

在本文的实现部分, 生成器采用 Transformer 编码器结构, 由若干 Transformer 层组成生成器和判别器的参数交替优化. 训练中, 首先, 固定生成器的参数, 通过最小化判别器的损失函数优化判别器权重. 然后, 固定判别器的参数, 通过最小化生成器的总损失函数优化生成器权重. 上述过程重复进行, 直至生成数据的频域分布$F(P_{\text{gen}})$与目标分布T_{freq}收敛.

2.3 生成解释

在对时间序列进行分段并生成频域扰动之后, 通过分析分段序列中的扰动影响, 揭示模型对序列的依赖性. 同时关注扰动对模型预测输出的直接影响, 以及扰动引起的时域和频域变化, 从多个维度量化每个分段序列的重要性. 具体过程如图 2 所示, 通过生成器生成扰动序列后输入需要解释的黑盒模型, 对比扰动前后模型输出差异以及原始序列和扰动序列差异进行计算得到可解释性结果.

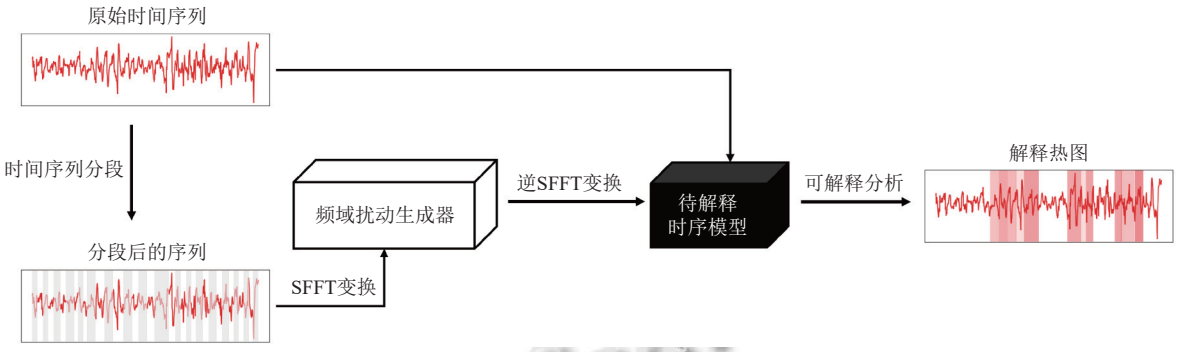

图 2 基于频域扰动的解释生成流程

首先, 对于每个分段序列S_i, GAN 根据目标频率分布生成对应的频域扰动δ_i. 这一扰动主要影响分段序列的特定频率成分. 在扰动生成后, 通过逆傅里叶变换将其从频域映射回时间域, 并叠加到原始分段序列S_i, 得到扰动后的序列S_i':

$$S_i' = S_i + \delta_i \tag{11}$$

随后, 将原始分段序列S_i和扰动后的分段序列S_i'分别输入到目标模型中, 记录对应的预测输出$\hat{y}_i$和$\hat{y}_i'$. 扰动引起的模型输出变化被定义为:

$$\Delta\hat{y}_i = |\hat{y}_i - \hat{y}_i'| \tag{12}$$

这一变化量直接反映了特定扰动对模型预测结果的影响, 输出变化越大, 表明该分段序列对模型预测的重要性越高.

为了进一步量化扰动对分段序列的整体影响, 引入时域变化量作为辅助指标. 扰动不仅会影响模型的预测结果, 也会改变分段序列本身的时域特性. 我们采用动态时间规整 (DTW) 算法计算扰动前后分段序列之间的形状变化, 用D_{DTW}表示:

$$D_{\text{DTW}} = DTW(S_i, S_i') \tag{13}$$

其中, DTW 能够捕捉分段序列的非线性形状变化, 较大的D_{DTW}值表明扰动对分段序列在时域上的影响较为显著.

此外, 我们还通过短时傅里叶变换 (STFT) 分析扰动对分段序列频域特性的影响. 通过提取扰动前后的频谱特性$P(f)$和$P'(f)$, 计算两者之间的频谱差异ρ_{freq}:

$$\rho_{\text{freq}} = \|P(f) - P'(f)\|^2 \tag{14}$$

频域变化量反映了扰动引起的频谱特性偏移程度, 较大的ρ_{freq}值表明扰动在目标频率范围内具有较强的作用.

结合上述 3 个指标, 最终定义分段序列的重要性指标$I_{imp,i}$, 以综合量化每个分段序列对模型预测的影响.

$$I_{imp,i} = w_1 \cdot \Delta\hat{y}_i + w_2 \cdot D_{\text{DTW}} + w_3 \cdot \rho_{\text{freq}} \tag{15}$$

其中, w_1, w_2, w_3为权重参数, 根据实验设置以平衡模型输出变化、时域变化和频域变化对分段重要性的贡献.

计算每个分段的重要性指标$I_{imp,i}$后, 为了直观地展示分析结果, 对每个分段的重要性指标进行可视化. 通过绘制时间序列的重要性热力图, 可以清晰展示每个分段在时间轴上的重要性分布.

3 实验分析

3.1 实验数据集及模型介绍

为了验证提出方法的性能, 我们在 AudioMNIST 和 UEA TSC Archive 选取的数据集上训练了时间序列分类模型并进行了可解释性分析实验.

AudioMNIST[25]数据集是一组用于语音分类和语音识别任务的公开数据集, 适用于研究语音生成的数字识别问题. 该数据集包含了数字 0–9 的语音录音, 由 60 位不同的说话者录制, 包括男声和女声, 涵盖多种音色和语调. 录音的采样率为 48 kHz, 每个音频片段的长度约为 1 s. 数据经过处理后, 可以转换为时间序列. 实验中对该数据集进行性别分类实验.

UEA TSC Archive (University of East Anglia time series classification archive)[26]是一个广泛用于时间序列分类研究的公开数据集集合. 该数据集包含来自多个领域的时间序列数据, 包括生物医学、气象学、金融、经济学、传感器数据、运动和环境等. 每个数据集都带有预定义的类别标签, 适用于时间序列的分类任务, 旨在推动时间序列分析技术, 特别是分类算法的研究和发展. 本文从中选取的 7 个数据集涵盖人体运动、手势识别等领域的时间序列数据, 具体如表 1 所示.

在实验中实现了一个基于 RNN 的时间序列分类模型, 具体来说, 使用了长短时记忆网络对整个时间序列进行编码, 并将最终时间步的隐藏状态传递到一个全连接层, 随后通过 Softmax 层生成每个类别的概率. 概率最高的类别被视为预测结果. 模型的具体参数如表 2 所示. 模型在 AudioMNIST 和 UEA TSC Archive 中选取的数据集上的分类性能如表 3 所示.

表 1 UEA TSC Archive 中选用的数据集

数据集	序列长度	类别数
AtrialFibrillation	144	25
FaceDetection	62	2
SpokenArabicDigits	93	2
Heartbeat	405	2
NATOPS	51	6
RacketSports	30	4
HandMovementDirection	400	4

表 2 实验训练的时间序列分类模型参数表

参数名称	值/说明
模型架构	RNN (基于 LSTM 单元)
LSTM 层数	1 层
隐藏状态维度	128
损失函数	交叉熵损失函数
优化器	Adam optimizer
学习率	0.001
Batch size	32
Epoch	500

表 3 时间序列分类模型分类效果

数据集	分类准确度 (%)
AudioMNIST	85.3
AtrialFibrillation	90.2
FaceDetection	66.5
SpokenArabicDigits	96.2
Heartbeat	74.2
NATOPS	88
RacketSports	77
HandMovementDirection	42

3.2 对比方法

实验中, 与基于梯度的方法, LIME 以及 LRP 方法进行了对比, 这些方法广泛应用于解释时间序列分类模型, 并且在许多研究中作为基准方法被使用.

基于梯度的方法[13]: 首先考虑的对比方法是 SG (saliency gradient), 通过直接计算预测类别的概率相对于输入特征的梯度来获得重要性分数. 此外, 还对比了 IG (integrated gradients), 通过沿着当前输入和用户定义的基线之间的路径数值积分梯度来计算重要性分数.

LIME[17]: 通过学习一个局部的可解释代理模型来计算特征的重要性分数. LIME 对输入数据进行扰动,

生成一系列人工数据, 并用原始分类器对这些人工数据进行预测. 接着, LIME 拟合一个简单的线性回归模型, 其中输入特征是表示原始特征“存在”或“缺失”的二进制向量, 输出是分类器的预测概率. 最后, 通过线性模型的权重绝对值来计算特征的重要性分数, 权重较大的特征被认为对预测结果影响较大.

LRP (layer-wise relevance propagation)[22]: LRP 方法通过反向传播的方式计算每个输入特征对模型预测的贡献. 具体来说, LRP 会从神经网络的输出层开始, 逐层传播预测的相关性. 每一层都会将总的相关性分配给前一层的神经元, 最终将预测的相关性分配到输入特征. LRP 通过逐层传播的方式, 帮助更好地理解神经网络内部各层的作用, 并提供输入特征对最终预测的贡献.

3.3 评价指标

在实验中, 我们通过计算 SDF (smallest destroying feature)[27]得分, 定量评估所采用的解释方法提供的解释. SDF 用于评估解释方法对模型预测的关键特征识别能力. SDF 通过逐步移除最重要的特征, 观察模型输出概率的下降速率, 衡量特征删除对模型预测的破坏程度, SDF 越低说明解释方法越准确地识别了不可或缺的关键特征.

3.4 结果与讨论

本节中, 我们将系统地评估提出的方法的性能. 首先, 对提出的扰动生成模型进行评估, 在图 3 中可视化了扰动生成模型生成的扰动序列以及与原始序列进行对比. 从图 3 可知, 所设计的频域扰动对时间序列的影响得到了有效的展示, 并且在不同频段上实现了较为显著的扰动. 具体来说, 图中的蓝色曲线代表扰动前的时间序列, 而红色曲线则展示了扰动后的序列. 我们在图中使用灰色区域标出了扰动作用的时段. 通过对高频和低频部分的特性分析, 可以发现, 在高频部分, 扰动引发了明显的波动增幅. 高频成分通常与快速变化的信号相关, 因此在这些部分能观察到时间序列的快速振荡和波动幅度的显著增加. 这表明, 频域扰动在高频区段的效应表现为对信号细节的加强或引入新的快速变化, 突出了扰动对短期动态特征的影响. 相较而言, 在低频部分, 扰动带来的变化虽然较为温和, 但依然可以观察到较为显著的特征改变. 低频成分通常与时间序列的长期趋势或周期性波动相关, 因此在这一部分, 扰动表现为对序列整体趋势的影响. 虽然低频部分的扰动幅度相对较小, 但依然能够体现出对长期趋势和全局行为的潜在影响.

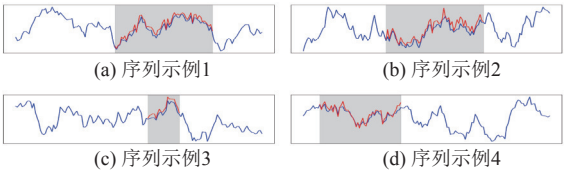

图 3 通过 GAN 生成的部分扰动序列

基于设计的频域扰动方法, 对多个数据集进行了可解释性分析实验. 在这些实验中, 我们选择了 SDF 作为主要的评价指标. 实验结果如表 4 所示, 最优结果使用加粗标明, 次优结果使用下划线标明. 并且我们可视化了部分可解释结果如图 4 所示, 其中红色部分越深表明该段序列的重要性得分越高.

表 4 本文可解释性方法与现有方法在 SDF 指标中的实验结果

数据集	IG	SG	LIME	LRP	本文方法
AudioMNIST	49.1	53.6	39.6	55.4	**28.8**
AtrialFibrillation	349.3	401.3	345.1	401.2	**301.2**
FaceDetection	124.0	**101.0**	145.3	143.1	108.9
SpokenArabicDigits	219.2	339.0	601.8	115.8	**87.1**
Heartbeat	**70.5**	76.5	109.4	82.5	**70.5**
NATOPS	446.9	656.4	822.5	**497.2**	498.6
RacketSports	64.0	96.4	57.3	**49.2**	52.9
HandMovementDirection	80.5	96.5	**77.8**	89.2	79.3

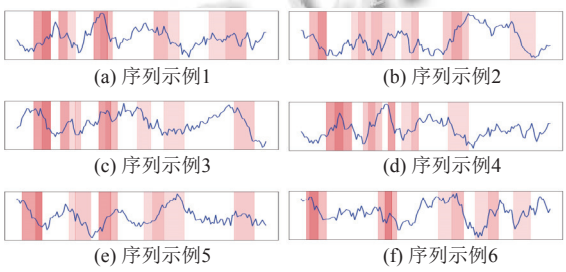

图 4 实验中部分序列解释热图

实验结果表明, 本方法在多个数据集上均取得了优秀的实验结果, 尤其在 AudioMNIST、AtrialFibrillation、SpokenArabicDigits 和 Heartbeat 这类与频率密切相关的数据集中, 表现尤为突出. 与现有方法相比, 本文方法在这些数据集上的效果远超其他方法, 展现了其在频域扰动和特征提取方面的显著优势. 这表明, 在处理与频率特征密切相关的数据时, 本文方法能够更有效地捕捉和扰动数据的频域信息, 从而实现更精确的扰动分析和模型解释.

此外, 实验结果还显示, 我们的方法在一些复杂的时间序列数据集上, 如 RacketSports、HandMovement-Direction 和 NATOPS, 也展现了较强的竞争力. 这些数据集涵盖了人体运动、手势识别等领域的复杂时间序列数据, 还涉及多种传感器数据的融合, 如加速度计、陀螺仪等, 反映了不同动作和状态下的动态变化. 在这些数据集中, 尽管信号的变化模式较为复杂且具有较强的时域特征, 我们的方法依然能够有效捕捉和扰动时间序列中的关键特征, 并且通过频域扰动, 不仅能有效提取这些时序数据中的频率成分, 还能够通过频域信息与时域特征相结合, 增强模型对这些复杂动态变化的敏感性. 这进一步证明了所提方法的广泛适用性, 不仅能处理传统上注重频率特征的数据集, 也能在时域特征较为复杂的数据集中发挥出色的作用, 有助于深入理解和优化模型的行为.

综合以上内容, 实验结果表明, 本文提出的频域扰动方法不仅在处理高频特征复杂的数据集时表现出色, 对于那些具有复杂时序动态的数据集, 我们的方法同样能够提供具有高解释性的扰动分析. 这表明, 本文提出的可解释性方法不仅能够有效捕捉和扰动频率特征, 还能够在时域特征复杂的场景中提供精确的分析, 帮助更好地理解和优化模型的行为.

4 结论与展望

本文深入研究了时间序列模型的可解释性问题, 尤其聚焦于复杂深度学习模型在处理时间序列数据时的“黑箱”特性. 为了解决这一挑战, 提出了一种创新的可解释性方法, 结合了频域扰动和时间序列分段算法. 通过引入基于频域分析的扰动生成机制, 不仅揭示了时间序列数据中不同频率成分对模型预测结果的影响, 还能够有效识别和分析模型依赖的关键频率特征.

具体而言, 通过短时傅里叶变换 (STFT) 和动态时间规整 (DTW) 算法, 将时间序列数据分段, 并利用生成对抗网络 (GAN) 生成频域扰动, 对每个分段序列的预测输出进行了详细分析. 实验结果表明, 所提方法在多个标准数据集上均取得了显著的效果, 特别是在处理高频特征较为复杂的任务中, 能够更精准地捕捉到时间序列中的频率变化, 对模型的预测结果提供深刻的洞察.

实验分析显示, 本文方法在提升模型透明度和可解释性方面表现出色, 相比于传统的解释方法, 如基于梯度的解释方法 (如 LIME 和 LRP), 在频率特征提取和扰动分析上展现了明显优势. 实验还表明, 在面对复杂动态变化的时序数据时, 频域扰动方法能够有效结合时域和频域特征, 为模型提供更为精确的可解释性分析.

未来的研究将进一步扩展该方法在多任务学习中的应用, 探索如何优化频域扰动的生成机制, 从而在不丧失模型精度的情况下, 提升其在复杂应用场景中的鲁棒性和解释能力.

参考文献

1 朱林. 一种基于金融时间序列数据的深度学习风险预测方法. 信息系统工程, 2024(6): 78–81. [doi: 10.3969/j.issn.1001-2362.2024.06.021]

2 屈峰. 基于时间序列分析的气象观测数据预测. 长江信息通信, 2024, 37(1): 36–39.

3 喻文杰, 陈宏文, 齐宏亮, 等. 基于长短期记忆网络和梯度提升的高血压患者 RR 间期时间序列预测方法. 中国医疗器械杂志, 2024, 48(4): 392–395. [doi: 10.12455/j.issn.1671-7104.230728]

4 王汝英, 马嘉骏, 董建强, 等. 基于 MTS-BiGRU-DMHSA 的工业负荷预测方法. 计算机工程, 2024, 50(9): 169–178.

5 张冬韵, 吴田军, 骆剑承, 等. 时空协同的农业种植结构遥感精细制图. 遥感学报, 2024, 28(8): 2014–2029.

6 谢丽霞, 王嘉敏, 杨宏宇, 等. 时间序列异常检测方法研究综述. 中国民航大学学报, 2024, 42(3): 1–12, 18. [doi: 10.3969/j.issn.1674-5590.2024.03.001]

7 王红军, 田铮. 非线性时间序列建模的混合自回归滑动平均模型. 控制理论与应用, 2005, 22(6): 875–881. [doi: 10.3969/j.issn.1000-8152.2005.06.006]

8 刘罗曼. 时间序列分析中指数平滑法的应用. 沈阳师范大学学报 (自然科学版), 2009, 27(4): 416–418.

9 玄英律, 万源, 陈嘉慧. 基于多尺度卷积和注意力机制的 LSTM 时间序列分类. 计算机应用, 2022, 42(8): 2343–2352. [doi: 10.11772/j.issn.1001-9081.2021061062]

10 郭丰玮, 王鹏新, 刘峻明, 等. 基于遥感多参数和 VMD-GRU 的冬小麦单产估测. 农业机械学报, 2024, 55(1): 164–174, 185. [doi: 10.6041/j.issn.1000-1298.2024.01.015]

11 任烈弘, 黄铝文, 田旭, 等. 基于 DFT 的频率敏感双分支 Transformer 多变量长时间序列预测方法. 计算机应用, 2024, 44(9): 2739–2746.

12 崔哲翰. 多元时间序列可解释性分类方法研究 [硕士学位论文]. 太原: 山西大学, 2024.

13 Sundararajan M, Taly A, Yan QQ. Axiomatic attribution for deep networks. Proceedings of the 34th International

Conference on Machine Learning. Sydney: JMLR.org, 2017. 3319–3328.
14 Modarressi A, Fayyaz M, Aghazadeh E, *et al*. DecompX: Explaining Transformers decisions by propagating token decomposition. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023. 2649–2664.
15 Crabbé J, Van Der Schaar M. Explaining time series predictions with dynamic masks. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 2166–2177.
16 Meng H, Wagner C, Triguero I. Explaining time series classifiers through meaningful perturbation and optimisation. Information Sciences, 2023, 645: 119334. [doi: 10.1016/j.ins.2023.119334]
17 Mishra S, Sturm BL, Dixon S. Local interpretable model-agnostic explanations for music content analysis. Proceedings of the 18th International Society for Music Information Retrieval Conference. Suzhou: ISMIR, 2017. 537–543.
18 文琪, 彭宏. 小波变换的离群时序数据挖掘分析. 电子科技大学学报, 2005, 34(4): 556–558 . [doi: 10.3969/j.issn.1001-0548.2005.04.033]
19 魏池璇, 王志海, 原继东, 等. 时间序列可变尺度的时频特征求解及其分类. 软件学报, 2022, 33(12): 4411–4428. [doi: 10.13328/j.cnki.jos.006346]
20 Brüsch T, Wickstrøm KK, Schmidt MN, *et al*. FreqRISE: Explaining time series using frequency masking. arXiv: 2406.13584, 2024.
21 Vielhaben J, Lapuschkin S, Montavon G, *et al*. Explainable AI for time series via virtual inspection layers. Pattern Recognition, 2024, 150: 110309. [doi: 10.1016/j.patcog.2024.110309]
22 Montavon G, Binder A, Lapuschkin S, *et al*. Layer-wise relevance propagation: An overview. In: Samek W, Montavon G, Vedaldi A, *et al*. eds. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019. 193–209.
23 肖洒, 陈旭阳, 叶锦华, 等. 一种基于 DTW-DP-GMM 的工业机器人轨迹学习策略. 天津大学学报 (自然科学与工程技术版), 2025, 58(1): 68–80.
24 徐慧兵, 李道伦, 查文舒. 基于梯度惩罚 Wasserstein 生成对抗网络的数字岩心重建. 合肥工业大学学报 (自然科学版), 2024, 47(11): 1559–1563.
25 Becker S, Vielhaben J, Ackermann M, *et al*. AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. Journal of the Franklin Institute, 2024, 361(1): 418–428. [doi: 10.1016/j.jfranklin.2023.11.038]
26 Bagnall A, Dau HA, Lines J, *et al*. The UEA multivariate time series classification archive, 2018. arXiv:1811.00075, 2018.
27 梁华杰. 深度神经网络的可解释性算法研究 [硕士学位论文]. 深圳: 深圳大学, 2022.

(校对责编: 张重毅)