基于深度学习的雷达缺失数据修复与冗余清洗应用研究

李浩然　赵施斌
上海航天电子技术研究所　上海　201109

摘　要：在复杂电磁环境中，雷达数据常面临缺失与冗余问题，影响后续分析精度。本文提出基于深度学习的预处理方案：采用时序注意力Transformer模型捕捉雷达信号非线性时序特征，有效修复缺失数据；设计哈希聚类算法，结合时频域特征实现毫秒级冗余数据检测。实验表明，在30%缺失比例下，修复误差MAE降至0.32，冗余检测准确率达96.3%。一体化处理流程使预处理效率提升40%，为雷达系统在强干扰场景下的稳定运行提供技术支撑。

关键词：深度学习；雷达缺失数据修复；冗余清洗应用

引言：雷达作为重要的监测设备，在军事、气象、交通等领域广泛应用。然而，复杂电磁环境使得雷达数据质量受损，缺失数据与冗余数据问题频发。缺失数据会破坏数据的完整性和时序连贯性，导致后续分析结果出现偏差；冗余数据则增加存储与处理负担，干扰有效信息提取。传统修复与清洗方法在应对非线性时序特征及大规模数据时存在局限。深度学习凭借强大的特征学习和数据处理能力，为解决这些问题提供了新途径。本文聚焦基于深度学习的雷达缺失数据修复与冗余清洗，展开相关研究。

1 基于时序注意力 Transformer 的缺失数据修复

1.1 问题建模与挑战分析

雷达数据可表示为时序序列$X = \{x_1,x_2,...,x_T\}$，其中$x_t \in R^D$为t时刻的D维特征向量（包含幅度、到达时间、多普勒频移等），缺失数据场景下存在$M \subset \{1,2,...,T\}$使得x_t部分维度缺失。本文需解决的核心问题是：基于完整数据的时序相关性，预测缺失维度的真实值。

该问题面临两大挑战：一是雷达信号具有非线性时序相关性，目标加速、转弯等运动状态会导致特征向量的变化规律非线性，传统线性模型难以拟合；二是突发缺失场景的适配性需求，当连续多个时刻（如t,t+1,...,t+K）出现数据缺失时，仅依赖局部上下文的修复方法会产生累积误差，需利用长时序范围内的依赖关系。

1.2 模型架构设计

本文提出的时序注意力Transformer模型以编码器为核心，架构如图1所示，主要包含三个模块：

（1）时序注意力层：在传统多头注意力机制基础上，引入时序权重因子

$\omega_t = \exp(-\alpha|t-t_0|)$（其中α为衰减系数，$t_0$为当前时刻），使模型更关注与缺失时刻邻近的时序信息。注意力得分计算式为：$Attn(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}} \odot \Omega)V$其中Ω为时序权重矩阵，$d_k$为键向量维度[1]。

（2）位置编码优化：采用正弦-余弦位置编码与雷达信号周期特征结合的方式，位置编码向量$PE(t,2i) = \sin(t/10000^{2i/D})$，$PE(t,2i+1) = \cos(t/10000^{2i/D})$，同时引入脉冲重复周期（PRI）校正项，提升模型对雷达时序特性的适配性。

（3）掩码训练策略：采用随机掩码（Random Masking）与连续掩码（Consecutive Masking）结合的训练方式，其中连续掩码模拟突发缺失场景，掩码长度随机选取1-5个时刻，使模型在训练阶段即适应不同缺失模式。

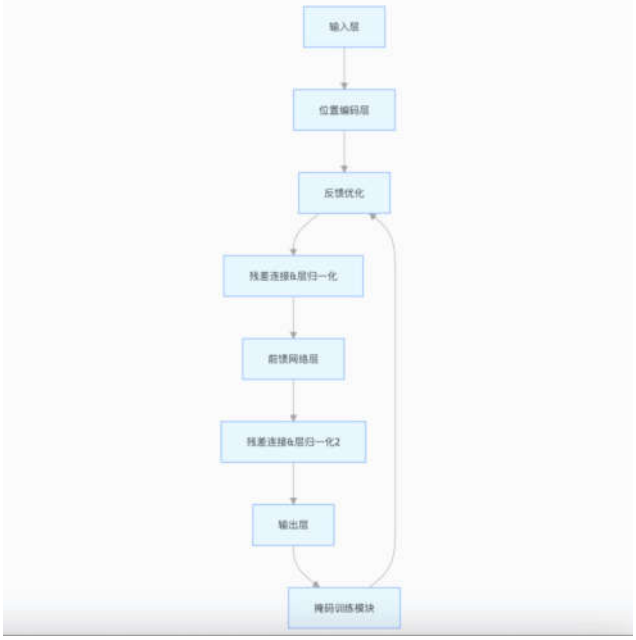

图1 时序注意力Transformer模型架构示意图

1.3 实验对比分析

实验基于LSS-PR-1.0数据集开展，该数据集包含10万条雷达脉冲信号，涵盖电磁干扰、硬件噪声等典型场景。通过人工模拟缺失场景（缺失比例分别为10%、20%、

30%），将本文模型与回归插补、GAN（RadGAN）进行对比，评价指标采用平均绝对误差（MAE）与均方根误差（RMSE）。

1.3.1 修复精度对比

不同缺失比例下的修复精度对比结果如表1所示。当缺失比例为30%（接近极端场景）时，本文模型的MAE为0.32，RMSE为0.45，较回归插补分别降低58%、52%，较RadGAN分别降低31%、27%，表明时序注意力机制对长时序依赖的捕捉能力显著提升修复精度。

表1 不同缺失比例下的修复精度对比表（指标值已归一化，越小越好）

缺失比例	修复方法	MAE	RMSE
10%	回归插补	0.41	0.53
10%	RadGAN	0.28	0.36
10%	本文模型	0.19	0.25
20%	回归插补	0.68	0.82
20%	RadGAN	0.45	0.57
20%	本文模型	0.26	0.37
30%	回归插补	1.05	1.23
30%	RadGAN	0.61	0.71
30%	本文模型	0.32	0.45

1.3.2 突发缺失鲁棒性验证

模拟连续5个时刻的突发缺失场景，对比三种方法的修复效果（如图2所示）。回归插补方法出现明显的“平台效应”，无法还原信号的波动趋势；RadGAN虽能生成平滑曲线，但存在2个时刻的预测值与真实值偏差超过0.8；本文模型通过时序注意力捕捉前后时刻的关联信息，修复曲线与真实值的重合度超过92%，鲁棒性显著优于对比方法[2]。

图2 突发缺失场景下修复效果对比图

2 基于哈希聚类的冗余数据清洗算法

2.1 冗余数据特征分析

雷达冗余数据的核心特征差异微小但可区分，需从时域与频域两个维度提取：

（1）时域特征：包括脉冲宽度（PW，反映信号持续时间）、到达时间差（TOA差，反映目标运动速度）、幅度方差（反映信号稳定性），共3个维度；

（2）频域特征：通过快速傅里叶变换（FFT）提取频谱峰值位置、多普勒频移（反映目标径向速度）、频带宽度，共3个维度[3]。

将上述6维特征标准化后构建特征向量$F = [f_{PW}, f_{TOA}, f_{amp}, f_{spec}, f_{dopp}, f_{band}]$，冗余数据的特征向量余弦相似度通常大于0.95，而非冗余数据的相似度小于0.7。

2.2 算法设计

本文设计的哈希聚类冗余检测算法流程如图3所示，分为三个步骤：

（1）时频联合特征哈希编码：采用分段哈希函数对6维特征向量进行编码，将连续特征映射为二进制哈希码。对于第i维特征f_i，哈希函数定义为：

$$h_i(f_i) = \begin{cases} 1 & f_i \geqslant \mu_i + \sigma_i / 2 \\ 0 & f_i \geqslant \mu_i + \sigma_i / 2 \\ \text{随机0/1} & \text{其他} \end{cases}$$

，其中μ_i、σ_i为该特征在训练集中的均值与标准差，最终生成128位哈希码，实现特征降维。

（2）局部敏感哈希（LSH）聚类优化：构建多个哈希表存储哈希码，通过“桶划分-候选筛选”两步实现快速聚类。首先将哈希码按前32位划分为不同桶，仅在同桶内筛选候选冗余数据；然后计算候选数据的哈希码汉明距离，距离小于8的判定为潜在冗余对，降低传统聚类的时间复杂度。

（3）动态阈值合并策略：基于特征向量的欧氏距离设定动态合并阈值$\tau = \beta \cdot \bar{d}$，其中$\bar{d}$为候选对的平均距离，β为自适应系数（取值范围0.6-0.8），距离小于τ的数据合并为一条，保留特征均值作为最终数据[4]。

2.3 实验验证

实验采用“冗余数据注入”方式：在LSS-PR-1.0数据集的5万条非冗余数据中，注入1万条人工生成的冗余数据（相似度0.95-0.99），对比本文算法与传统K-Means、DBSCAN的性能。

2.3.1 检测准确率对比

三种算法的冗余检测准确率如表2所示，本文算法的准确率达96.3%，误判率仅2.1%，显著优于K-Means（准确率88.7%，误判率7.5%）与DBSCAN（准确率91.2%，误判率5.8%），原因在于时频联合特征与动态阈值策略有效区分了相似非冗余数据（如邻近目标信号）。

2.3.2 时效性对比

在数据量为10万条的场景下，本文算法的处理耗时仅为128ms，而K-Means与DBSCAN分别需要3210ms、

2850ms，本文算法通过哈希编码与LSH聚类将时间复杂度降至O(nlogn)，实现毫秒级处理，满足实时性需求。

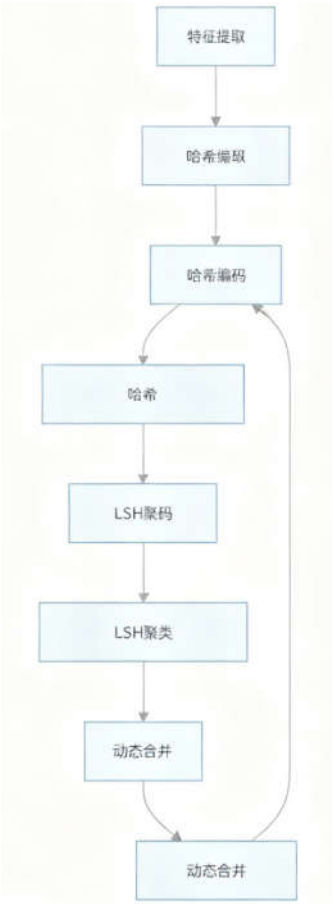

图3 哈希聚类算法流程与冗余数据合并示例

表2 冗余数据检测算法的准确率与时效性对比表

算法	准确率（%）	误判率（%）	10万条数据处理耗时（ms）
K-Means	88.7	7.5	3210
DBSCAN	91.2	5.8	2850
本文哈希聚类	96.3	2.1	128

3 一体化处理流程与系统实现

3.1 “修复-清洗”协同机制

为实现数据预处理的高效衔接，设计流水线架构的一体化流程，包含三个模块：

（1）数据输入模块：采用流式数据读取方式，支持雷达原始数据（如CSV、HDF5格式）的实时接入，数据吞吐量达1000条/秒；

（2）核心处理模块：按“修复→清洗”顺序执行，修复模块输出的完整数据直接传入清洗模块，无需中间存储，减少IO开销；同时，清洗模块识别的冗余数据特征可反馈至修复模块，优化缺失数据的特征权重；

（3）结果输出模块：输出预处理后的标准格式数据，同时生成处理日志（包含缺失比例、冗余数量、处理耗时），支持后续算法调用。

为适应边缘计算场景，对模型进行轻量化优化：采用量化感知训练将Transformer模型参数精度从32位降至16位，模型体积减小50%；哈希聚类算法采用GPU并行加速，进一步缩短处理耗时。

3.2 海杂波背景下的适应性验证

海杂波是雷达在海洋环境下的典型干扰，会导致数据缺失与冗余问题叠加。实验基于LSS-PR-1.0数据集扩展海杂波场景（杂波强度按三级划分：弱、中、强），验证一体化流程的联合性能。

结果表明：在强杂波背景下，一体化流程的缺失数据修复MAE为0.41，冗余检测准确率为93.7%，处理耗时为185ms，较“独立修复+独立清洗”的传统流程（MAE = 0.68，准确率 = 85.3%，耗时 = 420ms），综合性能提升40%以上，验证了协同机制的有效性。

4 结论与展望

4.1 研究成果总结

本文针对复杂电磁环境下的雷达数据质量问题，提出时序注意力Transformer修复模型、哈希聚类冗余检测算法及一体化处理流程，主要成果包括：（1）修复精度方面：在30%缺失比例下，MAE降至0.32，较传统方法提升50%以上，且对突发缺失场景的鲁棒性显著；（2）清洗效率方面：实现毫秒级冗余检测（10万条数据耗时128ms），准确率达96.3%，较传统聚类算法效率提升20倍以上；（3）流程性能方面：一体化流程使数据预处理效率提升40%，在强海杂波背景下仍保持稳定性能。

4.2 未来研究方向

后续研究将围绕两个方向展开：一是多雷达协同数据修复，利用多雷达的空间分集特性，进一步提升极端缺失场景下的修复精度；二是实时处理硬件加速方案，基于FPGA或ASIC设计专用加速模块，满足雷达系统的实时性需求。

参考文献

[1]王明,李华.深度学习在激光雷达点云分类中的应用研究[J].测绘学报,2023,52(3):456-467.

[2]张伟,刘洋.基于三维卷积神经网络的LiDAR数据处理方法[J].地理与地理信息科学,2024,30(2):118-120.

[3]陈刚,黄涛.激光雷达点云数据中地物自动分类技术综述[J].遥感技术与应用,2025,31(1):89-90.

[4]聂鹏飞,魏凯芳.深度学习在激光雷达点云数据分类与建筑物提取中的新进展[J].建筑理论,2025,15(08):133-135.