



北京工业大学
BEIJING UNIVERSITY OF TECHNOLOGY

深度学习在时序数据插补中的应用

汇报人：闫林枝

为什么要插补

➤ 数据缺失的现状^[1]:

传感器采集、船舶排放清单计算过程中产生缺失值、不规则时间间隔;

多变量时序特征易被破坏,影响后续建模和预测;

➤ 数据插补的意义:

恢复数据完整性,保持多变量关联;

减少模型偏差,提高预测精度

➤ 针对我的数据集:

CO₂、CH₄、N₂O多变量;

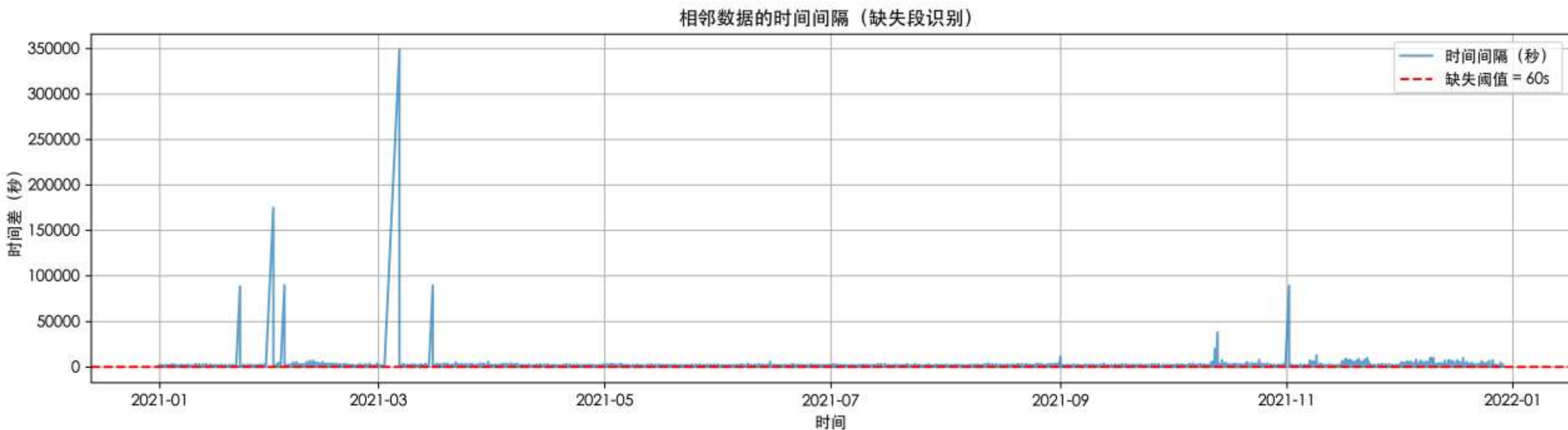
不规则采样+缺失率高;

后续要进行深度预测与减排分析,必须先插补。

✓ **结论:** 对比传统线性方法与深度学习方法,探索更优的数据插补模型。

数据集介绍

- 本研究使用的数据集包含 CO_2 、 CH_4 、 N_2O 三种污染物，采集于船舶排放清单，时间间隔不规律，有秒级、分钟级甚至小时级，同时缺失率较高。这类数据对插补方法提出了较高要求：既要处理多变量耦合，也要应对长缺口和不规则采样。



CONTENT_

01 RNN系列_

02 注意力机制_

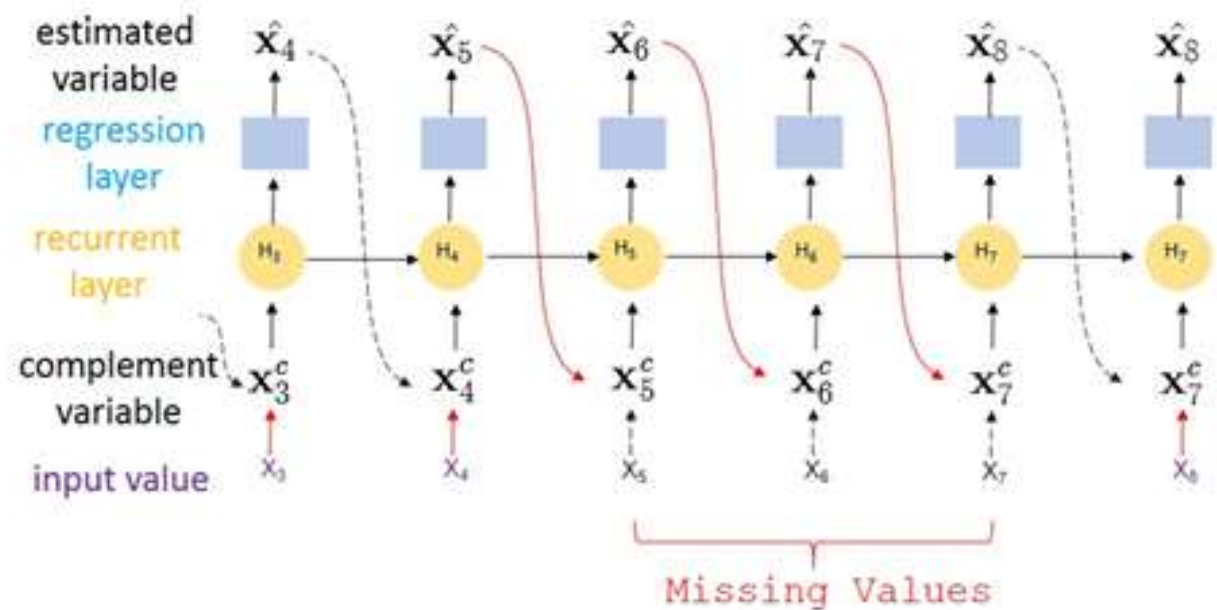
03 GAN_

04 结论_

RNN类插补模型

	主要特点	优点	缺点
BRITS ^[2]	双向RNN+时间衰减+特征相关性建模；把缺失值作为图中可学习变量；联合进行插补与分类/回归	无需对数据生成过程作线性或平稳性假设；前后向同时利用，插补精度高；多任务联合训练，有利于下游任务	模型结构较复杂，训练时间较长；需要较多计算资源；对小样本数据不够友好
MRNN ^[3]	双向RNN估算缺失值，但将插补值视为常量，不回传梯度；不显式建模特征间相关性	比单向RNN精度高；利用前后向信息；实现相对简单	插补值不能充分更新，精度不如BRITS；没有利用特征间相关性；下游任务效果一般
GRU-D ^[4]	基于GRU，引入时间衰减因子 γ 和全局均值，用上一观测值和均值平滑估算缺失值	模型简单、参数少、易训练；在医疗数据等平滑场景表现好；分类准确率高	插补是隐式的，插补精度差；只适合缺失模式较规律、平稳的数据；对多变量关联捕捉能力弱

BRITS



- RITS- I: 估计缺失值的误差会一直延迟到下一个观测值出现，模型收敛缓慢，训练效率低下。

- BRITS- I: 针对 RITS-I 中误差延迟导致模型收敛慢和偏差爆炸的问题，BRITS-I 利用双向循环，让每个值不仅能从正向推导，还能从反向推导。分别在正向和反向执行 RITS-I 算法，通过引入“一致性损失”，让预测在两个方向保持一致，最终的估计值是正向和反向估计值的均值。
- 相关循环插补：观测的特征不都是相互独立的，对于相关的特征，RITS利用特征相关性，除了基于历史数据的估计，还引入基于特征的估计，将两者结合得到最终值，BRITS是在双向情况下利用特征相关性进行计算。

注意力机制

	主要特点	优点	缺点
SAITS ^[5]	基于Transformer的双路径插补模型；采用双DMSA模块+加权组合模块；显式使用缺失掩码和位置编码	显式建模缺失值，插补精度高；同时捕捉时间依赖和特征依赖；可并行计算、训练速度快；	模型结构复杂、参数多；对小样本数据不够友好；需要较多显存和算力
Transformer ^[6]	标准自注意力机制，通过位置编码处理时间信息；单路径建模序列依赖	并行能力强、可捕捉长程依赖；结构通用，易与其他模块结合；训练稳定性较好	插补需额外改造；位置编码在非均匀采样场景下适应性差；参数量大
TimesNet ^[7]	利用多尺度卷积与傅里叶变换构建时序块，将时间序列映射到多时间尺度的频域；显式建模周期与趋势	更擅长提取周期性和多尺度特征；在长序列预测/建模上优于传统Transformer；计算效率高于纯Attention	主要针对预测任务设计，插补任务需额外适配；实现复杂度高

SAITS

两个任务

- ORT: 观测重建任务;
- MIT: 掩码插补任务;
- DiagMaskedMHA: 对角掩码自注意力机制, 应用这个对角掩码, 第 t 个时间步的输入值无法看到自身, 聚焦的是其他 $t-1$ 个时间步的输入值, 能够仅通过一次注意力操作就捕捉到高维空间中各时间步之间的时间依赖关系以及特征相关性。
- 第一个DMSA块: 输入是实际的输入特征向量和缺失掩码, 经过投影、位置编码等输出;
- 第二个DMSA块: 输入是第一个DMSA块的输出以及掩码, 最后有一个加权组合模型。

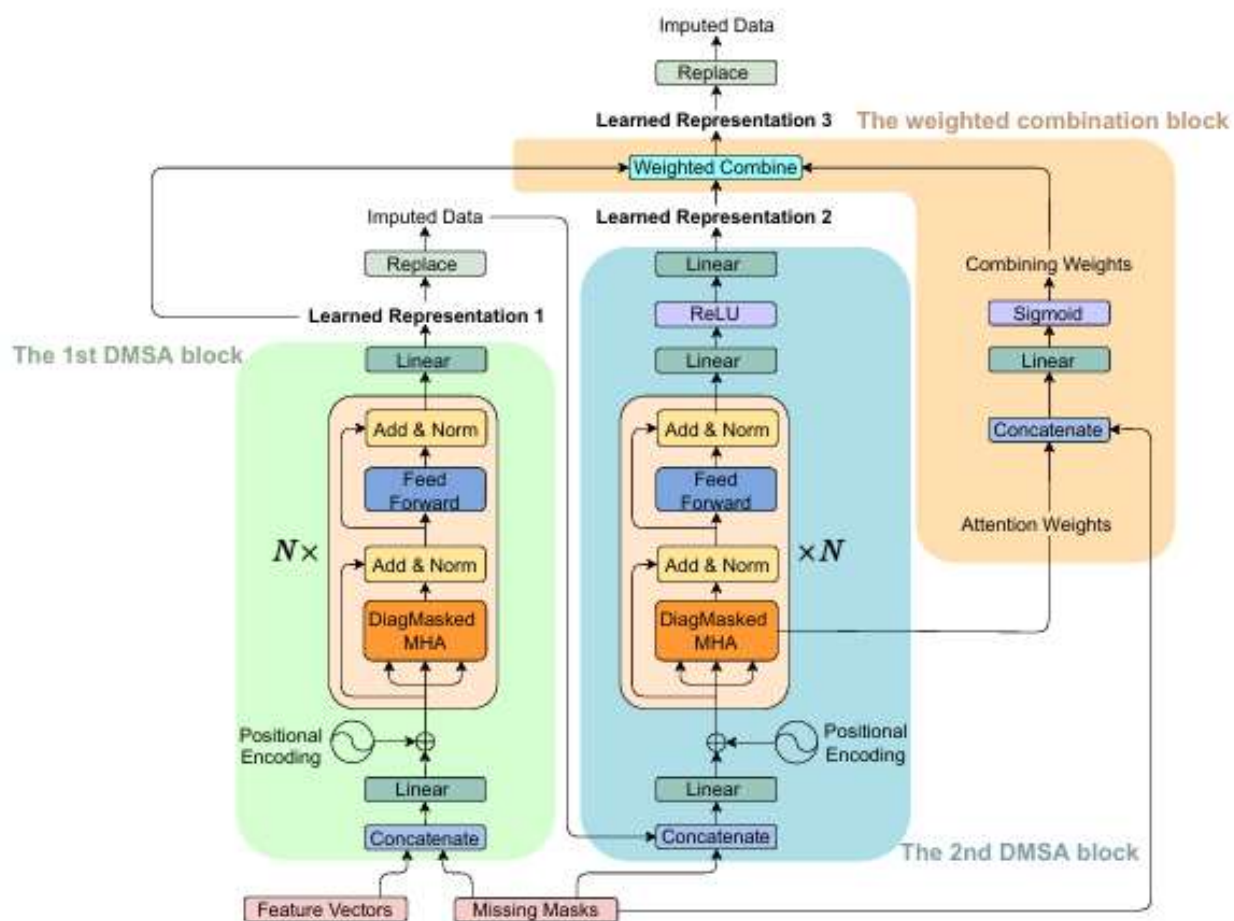
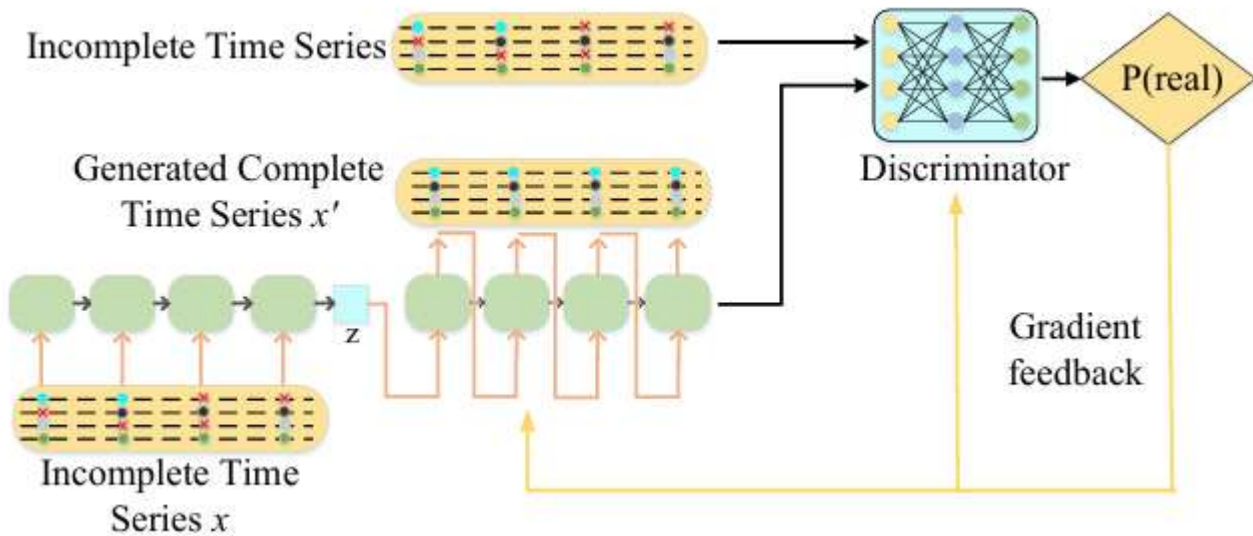


Figure 3: The SAITS model architecture.

GAN类插补模型

	主要特点	优点	缺点
E2GAN ^[8]	基于GAN的端到端时间序列插补模型；包括编码器、生成器、判别器以及循环单元来捕捉时序依赖	专门针对时间序列缺失值设计；能同时利用时序模式和全局分布特性；插补结果平滑、上下文一致性好	模型结构复杂、训练不稳定（GAN对抗训练常见问题）；需要较大数据量和算力；超参数敏感
GAIN ^[9]	在生成器中引入“Hint”机制指导判别器，主要用于表格数据插补	不需要强假设即可处理随机缺失；插补效果好、泛化能力强；可扩展到多种类型缺失模式	原始设计面向静态表格数据，时间序列需额外改造；对缺失率高的时序数据难以捕捉动态模式
NAOMI ^[10]	采用多分辨率、非自回归策略逐步从粗到细预测缺失值；利用双向信息提升插补精度	多分辨率推断加速训练，插补结果更稳定；非自回归避免误差积累；双向上下文提升插补精度	结构相对复杂，调参难度大；多分辨率模块需要精心设计；主要聚焦于序列预测，对表格数据不一定适用

E2GAN



- 端到端框架：避免了两阶段GAN需要逐样本优化噪声的低效问题，训练和推理速度显著提升。
- **Generator**：采用带噪声的自编码器结构（输入为不完整序列+随机噪声 η ），编码压缩到低维向量 z ，再解码重建完整序列 x' ；
- **GRUI单元**：在生成器和判别器中使用改进的GRU单元，引入时间衰减因子 β 建模不规则采样间隔，适应真实时间序列的缺失模式E2GAN。
- **Discriminator**：基于GRUI，区分真实序列与生成的补全序列。联合损失：使用对抗损失（来自判别器）+重建损失（MSE），确保插补结果既合理又接近真实值

总结

- ✓ **RNN类**：善于捕捉时间依赖和短期动态，对缺失值敏感性高，但并行性差。
 - ✓ **自注意力类**：用自注意力机制同时捕捉时间与特征关系，并行性强、可解释性好，适合高维长序列。
 - ✓ **GAN类**：通过生成对抗学习直接生成完整序列，端到端高精度插补，同时提升下游任务表现。
-
- 系统对比三类主流模型，总结其插补思路和特点，为后续选择或改进模型提供参考。
 - 引导未来工作：考虑多任务联合、物理先验（如船舶或柴油机实际特性）、更高效的混合架构。

参考文献

- [1] Wang J, Du W, Yang Y, et al. Deep learning for multivariate time series imputation: A survey[J]. arXiv preprint arXiv:2402.04059, 2024.
- [2] Cao W, Wang D, Li J, et al. Brits: Bidirectional recurrent imputation for time series[J]. Advances in neural information processing systems, 2018, 31.
- [3] Luo Y, Cai X, Zhang Y, et al. Multivariate time series imputation with generative adversarial networks[J]. Advances in neural information processing systems, 2018, 31.
- [4].Che Z, Purushotham S, Cho K, et al. Recurrent neural networks for multivariate time series with missing values[J]. Scientific reports, 2018, 8(1): 6085.
- [5].Du W, Côté D, Liu Y. Saits: Self-attention-based imputation for time series[J].
- [6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [7] Wu H, Hu T, Liu Y, et al. Timesnet: Temporal 2d-variation modeling for general time series analysis[J]. arXiv preprint arXiv:2210.02186, 2022.
- [8] Qin R, Wang Y. ImputeGAN: Generative adversarial network for multivariate time series imputation[J]. Entropy, 2023, 25(1): 137.
- [9] Yoon J, Jordon J, Schaar M. Gain: Missing data imputation using generative adversarial nets[C]//International conference on machine learning. PMLR, 2018: 5689-5698.
- [10] Liu Y, Yu R, Zheng S, et al. Naomi: Non-autoregressive multiresolution sequence imputation[J]. Advances in neural information processing systems, 2019, 32.

感谢观看！

汇报人：闫林枝

> GOODBYE <