中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST

2024 大数据挑战赛

BIG DATA CHALLENGE

中国·鄂尔多斯 ORDOS, CHINA

全国高等学校计算机教育研究会

清华大学 Tsinghua University

NERCBDS 大数据系统软件国家工程研究中心
NATIONAL ENGINEERING RESEARCH CENTER FOR BIG DATA SOFTWARE

答辩队伍：

元胞自动机

01 团队背景和成员简介

02 整体设计

赛题理解：

在本次比赛中，需要根据3850个**全球**气象自动站点的**温度**与**风速**的**两年历史观测**，结合**ERA5数据集**给出的周围9个格点的4个协变量训练模型，实现对**中国**站点**未来72小时**、**1小时间隔**的**温度**和**风速**预报。

目标变量：

- 两米高度的温度值（℃）
- 两米高度的风速的绝对值（m/s）

ERA5协变量：

- 十米高度的纬向风速（m/s）
- 十米高度的经向风速（m/s）
- 两米高度的温度T2M（℃）
- 均一海平面气压（Pa）

训练所用数据为全球气象数据，且没有给出经纬度信息，而线上测试为中国气象数据，二者的数据模式具有差异。因此解决本题的关键是如何提高模型的**泛化能力**。

评价指标：

MSE 即均方误差（Mean Squared Error），是衡量回归模型预测精度的一种常用指标，它表示预测值与实际值之间差异的平方的平均值，可以反映出模型预测误差的大小，MSE的值**越小**，表示模型的预测越准确。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

其中：MSE 是均方误差，n 是样本数量，Y_i 是第 i 个实际观测值，$\hat{Y}_i$ 是第 i 个预测值。

由于风速和温度两个变量存在差异，为了保证指标的可比性，以**标准化后温度MSE的10倍与标准化后风速MSE相加**作为最终得分。

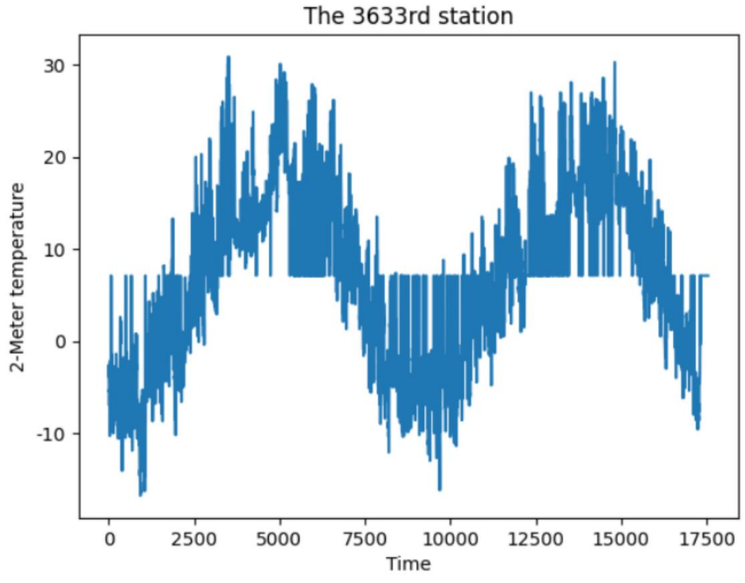
$$MSE_{\text{最终}} = \frac{MSE_{\text{风速}}}{Var\left(Y_{\text{测试集风速}}\right)} + 10\frac{MSE_{\text{温度}}}{Var\left(Y_{\text{测试集温度}}\right)}$$

总体框架：

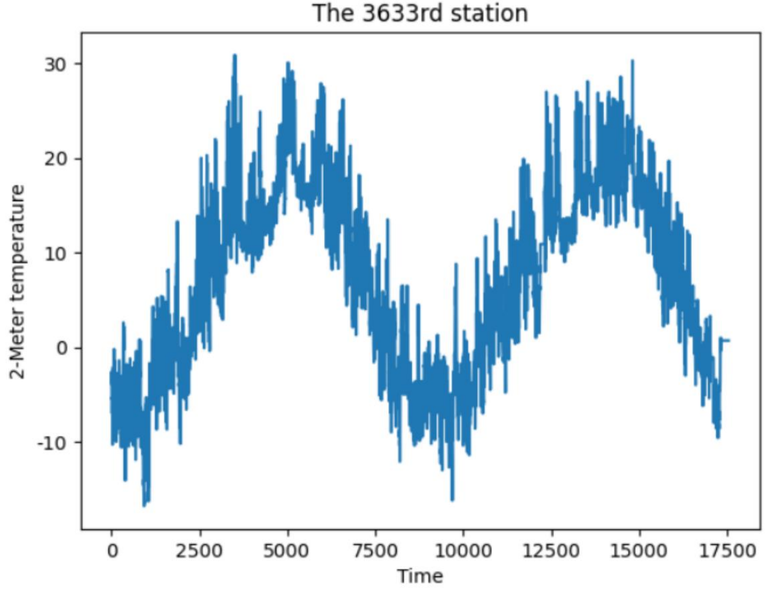
- 在探索性数据分析后进行**特征工程**，包括异常值处理、特征合并、构建新特征；
- 使用步长为1的滑动窗口构建数据集；
- 由于训练数据集为全球气象数据，测试数据为中国气象数据，二者数据模式不一致，所以在寻找最佳特征、模型结构和参数的时候使用**聚类交叉验证**——将数据经过**K-means聚类**后按簇交叉验证；**确定所有参数后**，使用**全量数据**训练模型；
- 模型采用**iTransformer+LSTM**架构；
- 使用多任务学习，一个模型**同时预测2米温度和2米风速**两个变量，将模型复杂度减半的同时充分利用二者之间的关联信息，有利于防止过拟合；
- 使用**MAE**和**Huber损失**，有效提高模型性能，并且在损失函数中**为后期时间步赋予更高的权重**；
- 最后，将10个不同参数不同损失函数训练得到的模型进行**集成预测。**

特征工程——异常值处理：

- 在数据可视化后发现**2米温度**数据中有大量**缺失值被用均值填充**，但是温度变量具有明显的**周期性**和**连续性**，这很可能使模型对数据模式产生错误的理解，因此修正为**使用前一个值填充缺失值**，从而保护了数据的周期性和连续性。



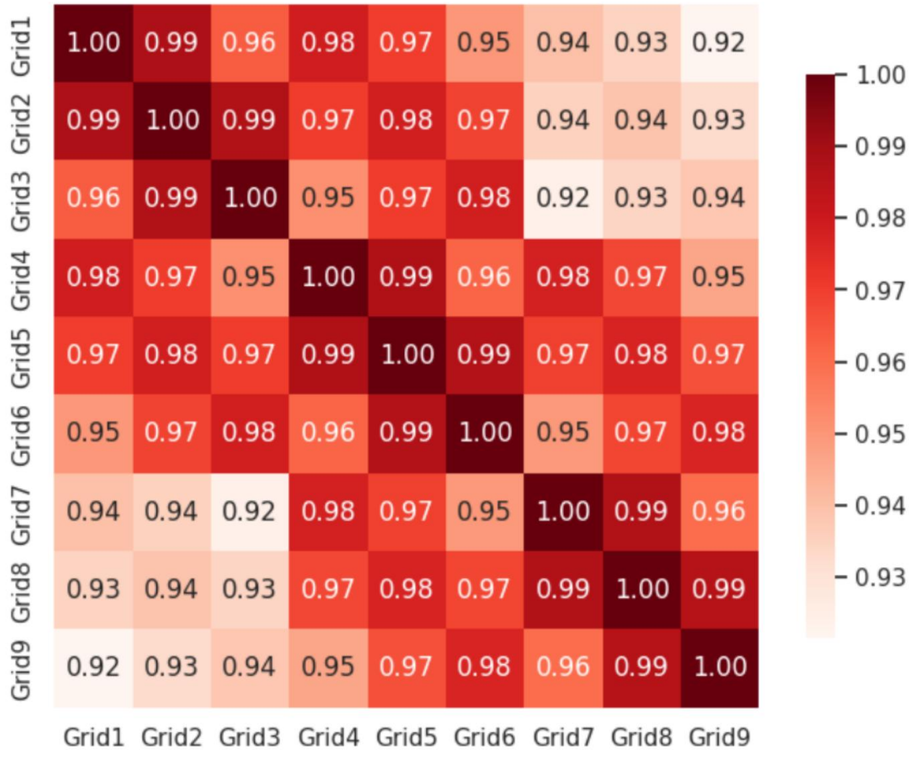
异常值处理前



异常值处理后

特征工程——特征合并：

- 对ERA5数据可视化后发现，**9个格点的数据具有非常明显的线性相关关系**，计算平均皮尔逊相关系数热力图见右图，相关系数最低的两个格点也达到了0.92，具有非常显著的线性相关关系；
- 因此**对9个格点的ERA5数据取均值**，使得Transformer模型能将更多的注意力放到除ERA5之外的特征上，大幅加快训练速度的同时提高了模型的性能。



ERA5数据9个格点间平均相关系数热力图

特征工程——建立新特征：

首先利用已有特征构建大量的**新特征**：

- 差分特征：2米风速的差分、2米温度的差分、均一海平面气压的差分、10米风速的差分、ERA5温度的差分
- 算数特征：矢量合成的10米风速、10米风速和2米风速的差、ERA5温度和2米温度的差、10米风向
- 交互特征：热通量Q、风冷指数WCI、2米风速和2米温度的积、均一海平面气压和10米风速的积
- 统计特征：均值、标准差、峰度

$$Q = V_{10m} \times (T_{ERA5} - T_{2m})$$

$$WCI = (10.45 + 10\sqrt{V_{2m}} - V_{2m}) \times (33 - T_{2m})$$

然后使用**前向选择法**和**聚类交叉验证**，选择**最佳特征**：

- 差分特征：2米风速的差分、2米温度的差分、均一海平面气压的差分
- 算数特征：矢量合成的10米风速、矢量合成的10米风速和2米风速的差、ERA5温度和2米温度的差
- 交互特征：热通量Q、风冷指数WCI。

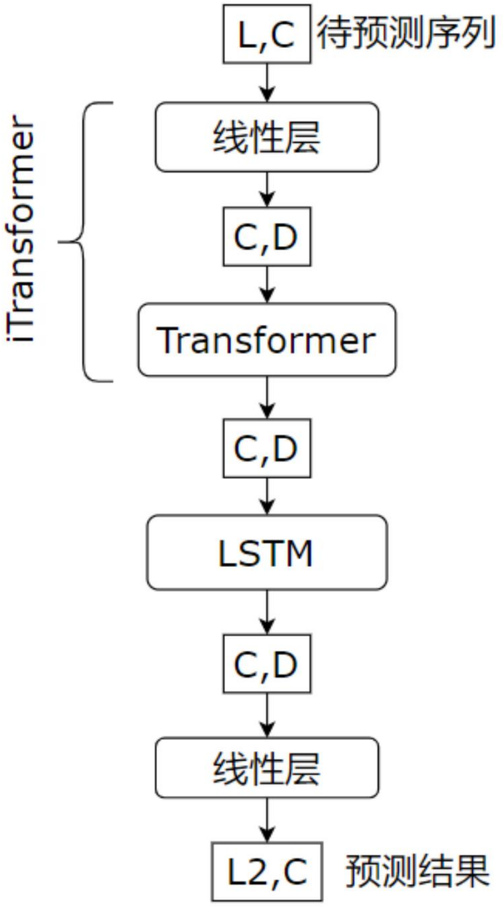
筛选出8个新特征，加上6个原特征，**共14个特征。**

模型结构：

- 使用清华大学提出的**iTransformer模型作为编码器**，iTransformer是一种倒置的Transformer结构，传统的Transformer将形状为(L,C)的序列编码为(L,D)进行学习（以时间为token）；而iTransformer将(L,C)的序列编码为(C,D)后再进行学习（以特征为token）。

- 使用一层**LSTM模型作为解码器**，iTransformer 在编码阶段能够捕捉特征之间的全局上下文信息，而 LSTM 则在解码时能够利用这些信息进行逐步预测，增强了模型的理解和预测能力。

- 右图为模型结构图，**待预测序列长度L**在本题中为168，**特征数C为14**，**维度D为超参数**，**预测序列长度L2**在本题中为72。



模型集成：

在训练和测试过程中发现，**单个模型的性能受超参数的影响较大**，因此使用十个不同参数的模型进行集成：

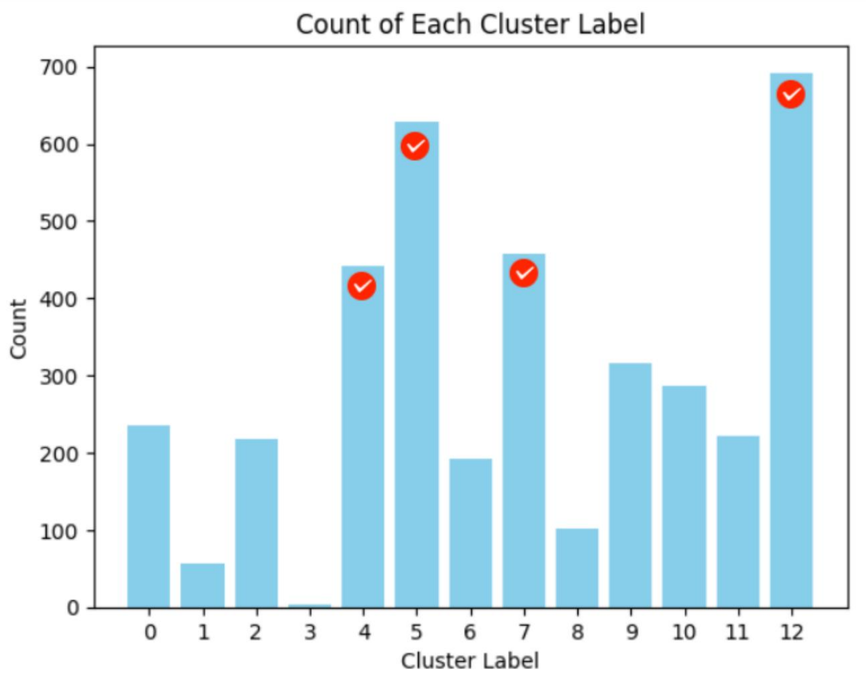
- 我们首先在控制变量的条件下开展大量实验，结果显示dmodel=256、lr=0.005、elayers=2时模型表现优秀且稳定易收敛；
- 然后对dff、nhead、loss三个超参数，使用**全量数据**训练十个异质的基模型，具体见下表，**其中*表示使用时间步加权损失；**
- 除此之外，为了提高单模型的**稳定性**和**抗噪性**，每个单模型使用**最后3个间隔800iter**的checkpoint和**最后3个间隔1000iter中损失最低的**checkpoint进行集成；

d_ff 前馈神经网络维度	512	512	512	512	1024	1024	512	512	512	512
n_head 注意力头数	4	4	4	4	4	4	8	8	8	8
损失函数	MAE	MAE*	Huber	Huber*	MAE	Huber	MAE	MAE*	Huber	Huber*

03 创新和实用

创新和实用——聚类交叉验证

- 数据分布：训练所用数据为全球气象数据，而线上测试为中国气象数据，二者的数据模式具有差异，这十分考验模型的**泛化能力**。
- 传统交叉验证的弊端：无论是时间交叉验证还是随机站点交叉验证，都会导致训练集和验证集之间的数据模式有较大重合，使用传统方法寻找出的特征、模型结构、参数无法保证泛化能力。
- 解决方法：使用**K-means聚类**方法**对站点进行聚类**，根据肘部法确定最佳K值在5~15之间，为了便于划分验证集，K值取13，聚类结果见右图，分别**选择数量占比前四的簇**作为验证集。
- 聚类得到的簇与其他簇在数据模式上有较大不同，使用这种方法进行交叉验证可以**很好地验证模型的泛化能力**。



创新和实用——模型结构

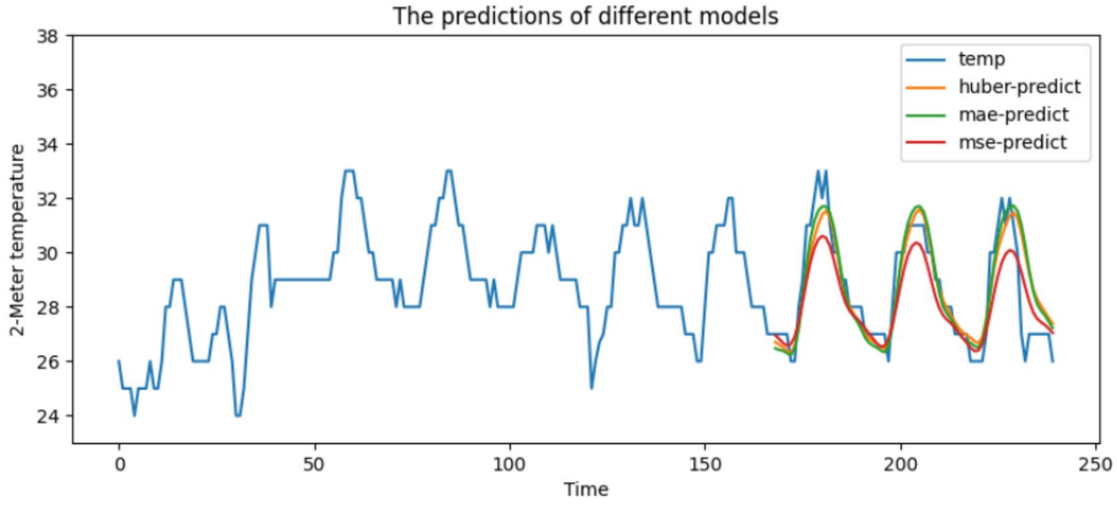
- 基础模型的选择：在初赛榜上对比iTransformer、SegRNN、PatchTST、TimesNet的性能后，发现iTransformer模型远远领先其他模型，于是固定使用iTransformer模型作为编码器
- 解码器的选择：在**固定超参数**（dmodels=256,dff=512,elayers=2,nheads=4,batchsize=15000,lr=0.005,loss=MAE）的情况下开展实验，寻找最佳的解码器，使用**聚类交叉验证**进行评估，结果见下表，可见选择一层LSTM模型作为解码器具有最好的泛化能力

	LSTM	2层LSTM	LSTM后不跟全连接层	无解码器	Transformer	GRU	CNN+LSTM
聚类交叉验证平均MSE	**1.23**	梯度爆炸	1.36	1.30	1.41	1.27	1.30

创新和实用——损失函数

- 在验证集上对预测结果可视化后发现：使用MSE作为损失函数时，模型的学习策略是过于保守的，其预测结果的**振幅往往偏小**，而MAE和Huber损失包含预测值与真实值之间的绝对差，**使得模型敢于给出大振幅的预测结果**，因此使用MAE和Huber作为基模型的损失函数。
- 为了进一步**增强模型长期预测的能力**，**使用时间步对损失进行加权**，从第一个时间步到最后一个时间步线性递增权重，最后一个时间步的权重是第一个时间步权重的二倍，并对权重进行归一化以保证loss的总和不变。
- 固定超参数进行聚类交叉验证结果见下图表，*表示加权损失

	MSE	MAE	Huber	MAE*	Huber*
聚类交叉验证平均MSE	1.293	1.232	1.235	**1.226**	**1.228**



创新和实用——训练优化和实用性

- 优化学习率调度策略：由于使用**步长为1的滑动窗口**构建训练集，导致**样本之间存在大量的重合。**如果每过1个epoch才调整一次学习率，会**导致大量算力被浪费。**我们将其优化为**每800iter调整一次学习率**，**训练2epoch**，batchsize为15000，**在1.2~1.5epoch模型即可收敛，大幅提高了训练速度。**
- 模型稳定性：由于使用了10个模型集成，每个模型包含6个checkpoint，大幅提高了集成模型的**泛化能力、稳定性和抗噪能力**，最终模型在**复赛B榜的MSE为1.1414**；在预测阶段对数据加入10%标准差的高斯噪声，MSE的变化保持在±0.0002以内。
- 超高的算法运行效率：使用2块V100训练全部模型，如果使用多线程训练，内存占用约40G，显存占用约45G，CPU使用率不超过1500%，**全部训练时间约为3小时**；**在测试时如果使用多线程推理，时间约为30秒**（含启动时间）。

04 总结与思考

总结与思考：

- ✓ 特征工程：异常值处理、特征合并、构建新特征；
- ✓ 模型架构：采用iTransformer结合LSTM，利用多任务学习降低复杂度并防止过拟合；
- ✓ 聚类交叉验证：使用K-means算法进行聚类交叉验证，提高模型的泛化能力；
- ✓ 损失函数：使用MAE和Huber损失函数，并对时间步进行加权，增强模型的预测能力；
- ✓ 学习率调度优化：增加学习率频率以大幅提升训练效率；
- ✓ 性能：最终模型在复赛B榜的MSE为1.1414，表现出色，全部训练时间仅需3个小时，算法运行效率高效。

- ➢ 未能充分利用ERA5数据集中9个格点的信息，可以考虑构建对角线格点经纬风速差的特征；
- ➢ 没有尝试使用数据的频域信息，如果尝试频域特征和频域损失可能有更好的效果提升；
- ➢ 未能充分探究单任务学习与多任务学习之间的差异，如果单任务学习并分开进行特征工程后，模型性能可能获得进一步提升。

中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST

感谢观看！