

降水量预测

自动化系 2019211242 王圣杰 2019211210 董宇光

摘要：

气象数据本质上是时间序列，机器学习算法具有挖掘其中序列特征的能力，适合利用大数据手段对该问题进行预测分析。但由于时间序列的随机性和随机性，设计一种可靠的计算方法是降雨预报中最具挑战性的任务之一。本文我们针对降水量预测问题，首先对降水量时序预测问题进行建模，利用皮尔逊系数得到对降水量相关系数较强的气象参数。然后利用 ARIMA、SVR、GBRT、XGBoost、FNN（MLP）、LSTM、seq2seq、atttention 机制的 seq2seq 模型对问题进行求解，探究了不同超参数下的寻优过程，实现了对下一个小时的降水预测，并在多个气象站得到了验证，并使用 RMSE、MAE、MDAE、R2-score 和可释方差指标对比了不同模型的预测效果。经过测试，单个模型中最为稳定的是 seq2seq 模型，经过模型集成中的 bagging 和 stacking 方法使得模型的预测性能进一步提高。此外我们还研究了不同输入序列长度对预测效果的影响，发现输入前三个小时的输入优于两个小时和一个小时，从 r2 得分上看提升了 20% 左右；在使用非时序的打乱的数据集训练和测试后，模型效果也进一步提升了 30% 左右；使用 atttention 机制的 seq2seq 模型来可视化不同地区所训练模型的特征重要性，通过分析得到了不同地区降水的特征信息。

关键词：时序预测 降水预报 机器学习 神经网络

1. 绪论

1.1 研究背景及现状

每年各类气象灾害给各国人民造成了巨大经济损失，也严重威胁着各国人民生命安全，其中和降水相关的干旱、暴雨、洪涝、泥石流等灾害又是最为常见的[1]。所以实现较好的降水量预测，对于防控各类气象灾害具有重大意义。

所以降水量预测任务是各国研究学者争相探索的领域，降雨是一个复杂的大气过程，具有时空依赖性，不易预测。由于降雨序列具有明显的随机性，它们通常被描述为一个随机过程[2]。有很多人试图找到最适合的降雨预报方法，例如，

将物理、海洋、气象或卫星数据与预报模型耦合，或者气象动力学结合统计学的方式来预测降水量。比如 Schepen 等利用 bayesian 模型平均法将动力和统计方法相结合，预测澳大利亚季节降水，结果显示模型优于单一动力学或统计模型[3]。还有一些利用时序分析的方法预测月降雨量序列或者年降雨量序列，如自回归模型（AR）[4]、分数高斯噪声模型[5]、自回归移动平均模型（ARMA）[6]和分解多元模型[7]。

近几年，机器学习在气象预测中的运用逐渐增加，在气象领域常见的机器学习算法有决策树模型，随机森林算法，人工神经网络，支持向量机等。Moustris 等人利用希腊四个气象站的长月降水时间序列，利用人工神经网络（ANNs）检验长期降水预报（连续四个月）的可能性[8]。还有一些研究学者，如 Ramana 等人利用小波技术和神经网络结合的模型运用到大吉岭雨量站的月雨量预测上，实现了比传统的神经网络和自回归滑动平均模型更好的效果[9]。

Bartoletti 等人基于神经模糊推理系统，运用数据驱动的方式建立降水量预测模型[10]。Cramer 等人研究了人工神经网络与遗传算法相结合的应用，以模拟 Teran 的月降雨量[11]。Lin 和 Jhong 利用多目标遗传算法开发了支持向量机（SVM）在台湾曾文河流域降雨小时预测中的应用[12]。Pham 等人研究了利用 ARIMA-MLP、ARIMA-LSSVM、ARIMA-NF 和 ARIMA-HW 混合模型对这些站的日降雨量进行了预测，对比单一模型，预测精度得到提高[13]。

除了降水量预测问题，环境中污染物预测也是近年来的热点，Zheng 等人基于大数据分析，建立未来 48 小时空气质量预测模型，该模型分为四个部分：1) 基于线性回归的时间预报器来模拟空气质量的局部因素；2) 基于神经网络的空间预报器来模拟全球因素；3) 结合根据气象数据预测的时空预报器的动态聚合器；4) 捕捉空气质量突然变化的拐点预测器。在对 48 个城市的的数据采集后，对模型进行评估，发现效果超过多个基线算法[14]。Liang 等人设计的 GeoMAN 网络用来预测多传感器时间序列问题，在空气质量数据和水质数据两种真实数据集上的实验表明，其方法优于九种基线方法，其方法主要采用了带注意力机制的 LSTM 网络来研究空间和时间上各类数据对最终效果的影响[15]。此外，在交通流量预测中，zhao 等人将图卷积网络和循环神经网络的变种 GRU 单元相结合，实现了从多个相关路段推测下一个时间段的交通流量信息，取得了较好的表现[16]。

1.2 研究内容

气象数据本质上是时间序列，机器学习算法具有挖掘其中序列特征的能力，适合利用大数据手段对该问题进行预测分析。但由于时间序列的随机性和随机性，设计一种可靠的计算方法是降雨预报中仍然是非常具有挑战的。本文我们的主要贡献在于 1) 针对降水量预测问题，首先对降水量时序预测问题进行建模，利用皮尔逊系数得到对降水量相关系数较强的气象参数。然后利用 ARIMA、SVR、GBRT、XGBoost、FNN、LSTM、seq2seq、Attention 机制的 seq2seq 模型对问题进行求解，探究了不同超参数下的寻优过程，实现了对下一个小时的降水预测，并在多个气象站得到了验证，并使用 RMSE、MAE、MDAE、R2-score 和可释方差指标对比了不同模型的预测效果。2) 使用模型集成中的 bagging 和 stacking 方法使得模型的预测性能进一步提高。3) 此外我们还研究了不同输入序列长度对预测效果的影响，发现输入前三个小时的输入优于两个小时和一个小时，从 r2 得分上看提升接近 20%。4) 在使用非时序的打乱的数据集训练和测试后，模型效果也进一步提升了 30% 左右。5) 使用 Attention 机制的 seq2seq 模型来可视化不同地区所训练模型的特征重要性，通过分析其中的区别为我们后续继续改善模型打下基础。

1.3 数学描述

本文提出的降水量预测的目标是根据该气象站历史的气象数据，对下一小时内降水量进行预测，气象数据为数据集中所代表的每个小时的各类气象参数。

首先，定义降水量预测的输入特征矩阵为 $X^{N \times P}$, 其中 N 为输入的前 N 小时数据，P 为对应的该小时的气象数据。降水量预测的输出为 Y，代表下一时刻的预测降水量。因此，降水量预测问题可以看作是在特征矩阵 X 的前提下学习映射函数 f，然后计算未来时刻的降水信息，如下式所示，其中 n 是历史时间序列的长度。

$$Y = f(X_{t-n}, \dots, X_{t-1}) \quad (1)$$

2. 数据整理

2.1 数据来源及内容

气象数据的种类多种多样，主要包括地面观测资料、天气雷达资料、气象卫星资料和数值预报产品。在本次使用的数据集上提供地面观测数据(数据集已上传至清华云盘链接)，数据集中各列的含义参见 abbreviated.txt。该数据集包含了

对 122 个气象站从 2000 年到 2016 年逐小时观测数据（并不是所有气象站都是从 2000 年开始观测的），包含 17 个气象参数。其中主要的 17 个气象参数为每小时降水量、气压、最大气压、最小气压、太阳辐射、空气温度、最高温度、最高露点温度、最低温度、最低露点温度、相对湿度、最高湿度、最低湿度、风速、风向、阵风。除此之外还记录了各地的海拔高度、经度、纬度。

2.2 数据清洗

在处理其中的数据时我们可以发现有些数据位为缺失值，观察后发现大部分缺失的情况是气象上该时段没有记录，所以将这些缺失值置为 0，可以充分利用数据集。但对于有些时间段，所有的气象数据均为 0 的情况，此时将该行数据删除，避免干扰运算。此外，由于我们采用的是每天气象的数据，所以如果数据集中所选用特征在该小时内均为 0 的点采用线性插值的方式解决缺失值。

2.3 可视化观察数据分布

选取了几个气象点的每小时降水数据作为观察，下图为其中的一个例子，可以看出降水量在小范围的时间段非线性关系很强，但在大的周期上看呈现一定的规律，比如一个周期内降水比较频繁。通过计算其他气象数据和降水量的皮尔逊系数，可以得到如下结果。可以看出其中太阳辐射值、露点温度、相对湿度还有阵风有较大的影响，所以选择其作为输入模型的特征，为了对比效果也引入温度、风向等参数输入模型。

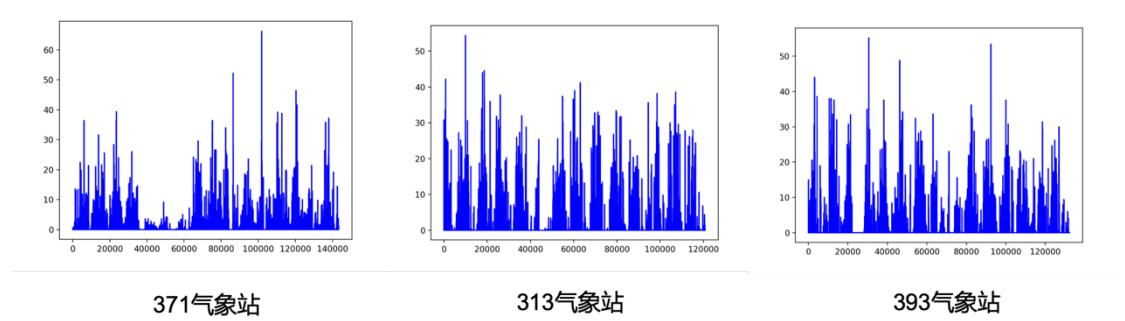


图 1 数据集可视化

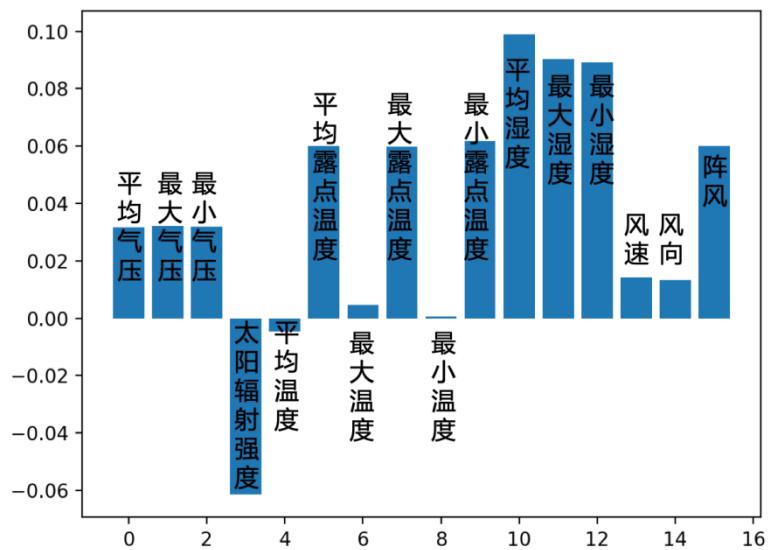


图 2 皮尔逊系数相关性检验

3. 特征与模型设计

3.1 特征选择

对于备选的特征，采用 0 均值和单位方差的归一化方式来消除数据中数值的影响。

对于特征间的组合，考虑到这是一个时序预测问题，此外在之前的文献中也发现在预测时研究者都会输入前 n 天的特征数据来做预测，在多篇文献中也探究了具体输入的时序长度对结果的影响。除此之外我们还测试了不同小时数对模型的影响，最终确定采用了前 3 个小时的气象数据来做预测。

我们用 o_t^i 代表 t 时刻的观测特征 i。其中 $i=1,2,3,4,5$ ，分别代表 t 时刻的降水、太阳辐射强度、平均温度、平均露点温度、平均湿度、风向、阵风速度。所以 t 时刻的观测特征为 $o_t = [o_t^1, o_t^2, o_t^3, o_t^4, o_t^5]$ 。定义预测 t 时刻的降水量模型的输入向量为 $x_t = [o_{t-2}, o_{t-1}]^T$ ，降水量的真实值为 y_t ，输出的降水量为 \hat{y}_t 。

3.2 模型设计

3.2.1 ARIMA

ARIMA 模型全称为自回归移动平均模型，是常见的一种用来进行时间序列预测的模型。其方法在于通过对序列差分获得平稳信号，之后可以根据下式的自回归移动平均模型进行拟合得到预测值。

$$\Delta^d y_t = \alpha_0 + \sum_{i=1}^p \beta_i \Delta^d y_{t-i} + \sum_{j=1}^q \alpha_j \varepsilon_{t-j} \quad (2)$$

其中 y_t 为训练集中降水量序列， $\Delta^d y_t$ 为 d 项差分后的平稳序列。 ε_{t-1} 为 0 均值的白噪误差序列。其中 α_j 和 β_i 为模型的估计参数， p 和 q 为模型的阶数。对于该模型在我们问题的应用，由于观察到降水序列一项差分后的结果就已经为平稳信号，应该主要关注 p 和 q 参数的取值。

3. 2. 2 SVR

SVR（支持向量回归）模型是一种从历史数据中拟合一个线形表达式的估计模型。其主要的目的在于找到一个回归平面，让一个集合的所有数据到该平面的距离最近。具体的求解思路类似 SVM 模型，SVR 在线性函数两侧制造了一个“间隔带”，对于所有落入到间隔带内的样本，都不计算损失；只有间隔带之外的，才计入损失函数。之后再通过最小化间隔带的宽度与总损失来最优化模型。

3. 2. 3 FNN

FNN 网络是从简单的单一感知器而发展而来，前馈神经网络主要由输入层，隐藏层和输出层组成。对于单层的前馈神经网络，其输入输出的变化关系为：

$$s_j = \sum_{i=1}^n w_{ji} x_i - \theta_j \quad (3)$$

$$y_j = f(s_j) \quad (4)$$

其中 x_i 代表输入特征的第 i 个分量， w_{ji} 代表 x_i 和 y_j 直接的连接权重， y_j 为对应的输出的第 j 个分量， θ_j 为对应的偏置量， f 为对应的激活函数。

3. 2. 4 LSTM

LSTM 网络为一种循环网络（RNN）的变形形式，全称为长短时记忆网络。其优秀之处在于解决了 RNN 网络中对长时间序列可能出现的梯度消失问题。具体的公式为如下表达式：

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (5)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (6)$$

$$g_t = \sigma(W_c x_t + U_c h_{t-1} + b_c) \quad (7)$$

$$c_t = i_t \odot g_t + f_t \odot c_{t-1} \quad (8)$$

$$output_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (9)$$

$$h_t = output_t \odot \tanh(c_t) \quad (10)$$

其中 f_t , i_t , g_t 分别为遗忘门, 输入门, 更新门来控制相关单元的输入, h_t 为隐藏的输出状态, $output_t$ 为输出, σ 代表的是 sigmoid 激活函数。

3.2.5 GBRT

GBRT(Gradient Boosting Decision Tree) 名为渐进梯度回归算法, 是一种迭代的决策树算法, 该算法由多棵决策树组成, 所有树的结论累加起来做最终答案, 其泛化能力较好。具体的思路为利用最速下降的近似方法, 即利用损失函数的负梯度在当前模型的值, 作为回归问题中提升树算法的残差的近似值(伪残差), 拟合一个回归树。算法的流程为 1) 初始化一个只有根节点的树, 2) 循环对第 m 个决策树的每个样本计算损失函数的负梯度, 3) 第 m 个回归树将空间分为网格性的小区域, 4) 重新计算每个区域的输出值, 5) 更新第 m 个回归树, 6) 循环所有的 M 个决策树后为最终的回归树。

3.2.6 XGBoost

XGBoost 算法在 GBRT 算法的基础上给代价函数加一个正则化项, 并且对目标函数进行二阶泰勒近似, 然后采用精确或近似方法贪心搜索出得分最高的切分点, 进行下一步切分并扩展叶节点。这样做, 一方面保证最小化损失函数的过程中, 树结构不会过于复杂而产生过拟合, 另一方面加快计算速度。此外, 叶节点输出值不再采用取叶节点包含样本均值的方式, 而是计算出每个叶节点最优权重。这样对于回归问题, 最终输入的各个叶节点的权重即为最终的预测值。

3.2.7 Seq2seq 网络

Seq2seq 网络模型最先用于自然语言翻译中, 它对于时序序列的预测效果显著, 所以针对我们的问题也可以采用了类似的模型解决。在 Seq2Seq 模型结构中, 由两个部分组成, 其一为编码器网络, 由多个 Rnn 网络连接构成, 其二为解码器网络, 同样由 Rnn 网络连接而成。整个网络流程为输入序列通过 Rnn 时序进入模型, 得到编码后的结果, 再输入解码器网络输出最终的时序序列。具体的结构如下图所示。

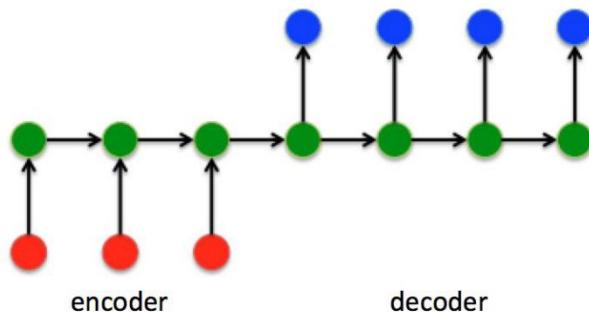


图 3 seq2seq 网络结构

此外，如果在解码网络中添加 Attention 机制，可以获得输入序列的各项重要性权重，这既有利于训练的过程，使得充分利用时序关系，也可以通过这个网络分析特征的重要性。具体的 Attention 机制不再过多解释，可以参考 Seq2Seq 的论文。

3. 2. 8 模型集成

Ensemble 名为模型集成，主要的算法有 Bagging, Boosting 和 Stacking。其中 Bagging 的算法思路为利用可放回抽样总共获得 N 个样本，之后训练 N 个分类器，对于回归问题取这 N 个分类器的均值，对于分类问题让这 N 个分类器进行投票，最终可以降低过拟合的程度。与 Bagging 的目的不同，Boosting 可以提高弱分类器的性能，如果使用的分类器在训练集上的误差小于 50%，通过 Boosting 之后分类器的错误率最终总会到 0。而最后一种 Stacking 的思路是将训练数据分割，用其中的一部分训练前几个模型的参数，然后再前几个模型的输出后接入一个简单的分类或回归模型，比如最小二乘或者决策树，用剩余的训练集训练最后的分类或回归模型，最终在测试集上取得优于单独模型的结果。本文我们主要采用了 Bagging 和 Stacking 这两种模型集成方法，通过实验分析得到了比单一模型更好的结果。

4. 实验设计

4. 1 数据集切分

通过预处理将小时记录的原始数据集通过计算得到每天的气象数据集。然后选取 7:2:1 比例的训练集、验证集和测试集。其中训练集验证集随机打乱，训练验证集和测试集按时间前后分开。

4. 2 评价指标

为了评估模型的预测效果，我们使用了 5 种度量方式来评估预测值 \hat{y}_t 和实际值 y_t 的误差。

1. 均方根误差：

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_t - \hat{y}_t)^2} \quad (11)$$

2. 平均绝对误差：

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (12)$$

3. 中值绝对误差:

$$MDAE = median(|y_1 - \hat{y}_1|, \dots, |y_n - \hat{y}_n|) \quad (13)$$

4. 决定系数:

$$R^2 = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y}_t)^2} \quad (14)$$

5. 解释方差得分:

$$var = 1 - \frac{Var\{Y - \hat{Y}\}}{Var\{Y\}} \quad (15)$$

具体来说，使用 RMSE、MAE 和 MDAE 来测量预测误差：预测误差值越小，预测效果越好。 R^2 和 Var 计算相关系数，衡量预测结果代表实际数据的能力：值越大，预测效果越好，实验中二者的值相差不大。

4.3 超参数寻优

这部分我们主要探究不同超参数对于模型泛化性能的影响，此处为了简化分析流程，举例 371 气象站来说明超参数寻找的过程，其余气象站均可以通过以下方法得到最优超参数。

首先对于 ARIMA 模型的超参数大部分是通过平稳性检验过程得到，本文不做过多的解释。超参数的研究主要针对 SVR、FNN、LSTM、SEQ2SEQ、GBRT、XGBoost 算法。其中 SVR 模型主要的超参数有：核函数 K、正则化系数 gamma、惩罚系数 C。为了选择最佳值，我们使用贪心寻优的方法先后对不同的参数进行调优实验，并通过比较预测验证集的 MSE 来选择最佳值。

4.3.1 GBRT、XGBoost

对于 GBRT 和 XGBoost 的算法的超参数主要是树的数量 m_{tree} ，最大深度 d_{tree} ，学习率 l_r 。从图中可以看出随着树的数量增大，训练 rmse 值下降，但验证集 mse 值先下降后上升。类似的结果同样从最大深度和学习率的曲线中得到。GBRT 的主要超参数是树的数量 m_{tree} ，最大深度 d_{tree} ，学习率 l_r 。这里寻优过程是先 lr ，再 est ，最后是 $depth$ ，分别代表学习率 l_r ，树的数量 m_{tree} ，最大深度 d_{tree} 。

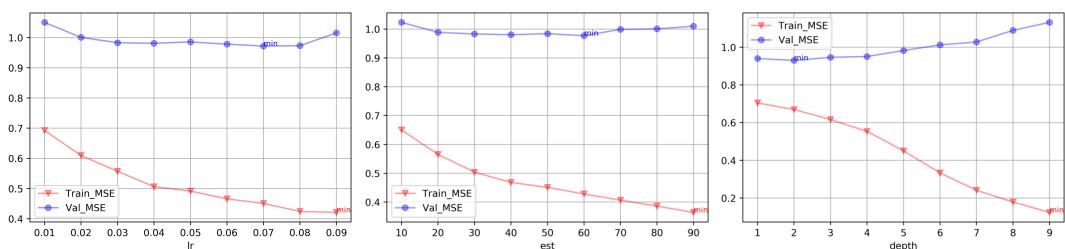


图 4 GBRT 的超参数寻优过程

在寻找最优学习率 lr 时，训练集损失随 lr 增大不断减小，在 0.09 取极小；验证集损失先减小趋于平缓，在 0.07 取极小，后有增大的趋势，于是便停止继续在 l_r 上寻优。开始在 est 超参数上寻优，与 lr 寻优结果类似，随着 est 增大，训练集损失不断减小，而验证集损失先减小后增大趋势，在 60 取极小。接下来在 depth 超参数下寻最优，在 depth 取 2 时验证集损失达到最优，而训练集损失也是一直在减小。因此这里最优超参数的组合为 lr = 0.07, est = 60, depth = 2。

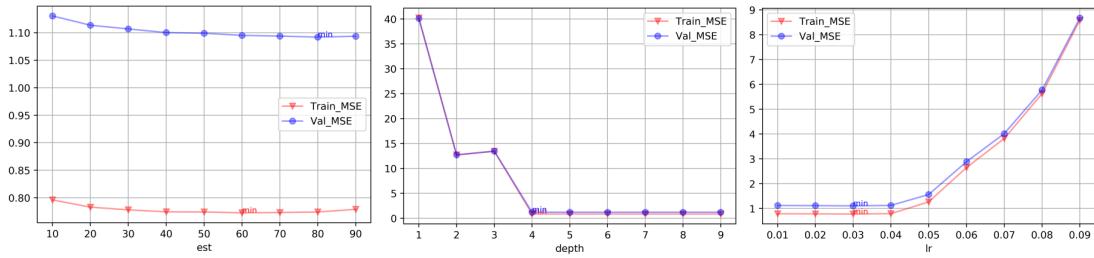


图 5 XGBoost 的超参数寻优过程

如图 5 是 XGBoost 算法的寻优过程。可以看到在训练集和验证集的损失上，二者变化是非常同步的，在 est=60, depth=4, lr=0.03 时二者均达到极小。

4.3.2 神经网络

对于 FNN (MLP)、LSTM 和 Seq2Seq 网络，三者均为神经网络的结构，所以他们的超参数类似，主要为隐藏单元个数。为了获得最优的预测效果，我们设计了不同的隐藏单元个数，并通过验证集的 MSE 值来取得最优模型。在我们的实验中，对于其中气象站 371 的降水量数据，我们从 [8, 16, 32, 64, 100, 128] 中选择隐藏单元的数目，并分析预测精度的变化。如下图所示，水平轴表示隐藏单元的数量，垂直轴表示不同度量的变化，其显示了不同隐藏单元的训练 MSE 和验证 MSE 结果。可以看出，在 FNN 中，当隐藏单元设置为 100 时，预测结果最好。当隐单元数增加时，预测精度先增大后减小。这主要是因为当隐藏单元大于一定程度时，模型的复杂度和计算难度大大增加，从而降低了预测精度。因此，我们在对其实验中将隐藏单元的数量设置为 100。同样的，LSTM 和 Seq2Seq 网络的最优隐藏单元为 100 和 100。

如图分别为 FNN、LSTM、Seq2Seq 网络寻优过程。

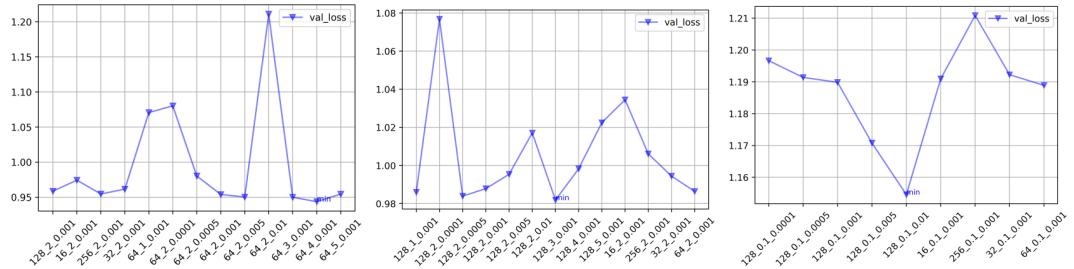


图 6 神经网络超参数寻优过程

从图中可以看出，FNN 在 64_4_0.001 超参数下最优，即隐藏层维数 64、隐藏层数 4、学习率 0.001；LSTM 在 128_3_0.001 超参数下最优，即隐藏层维数 128、递归网络个数 3、学习率 0.001；Seq2Seq 在 128_0.1_0.01 超参数下最优，即隐藏层维数为 128、学习率为 0.01（注：此处 0.1 是无效参数）。

4. 3. 3 SVR

下图为不同核函数（多项式核、高斯核、sigmoid 核）在正则化系数和惩罚系数的训练 mse 和验证 mse 的变化图。按顺序分别为多项式核、高斯核、sigmoid 核。

对于多项式核函数下的寻优，如下图。首先对于多项式核函数的超参数主要为核函数的次数 degree，惩罚参数 c 和核系数 gamma。按照重要性，我们首先寻优最优的核函数次数 degree，然后是惩罚系数 c，最后为核系数 gamma。可以看到 SVR-ploy 的模型效果不是很好，寻优中并未发现好的结果。

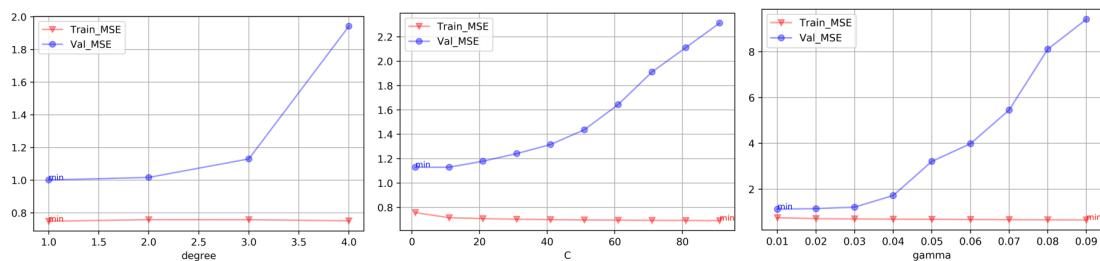


图 7 SVR-ploy 的超参数寻优过程

对于 sigmoid 核函数下的寻优，和多项式核函数一样，但其超参数只有惩罚参数 c 和核系数 gamma。同上述方法，我们发现结果也不是很好，甚至对超参数异常敏感。

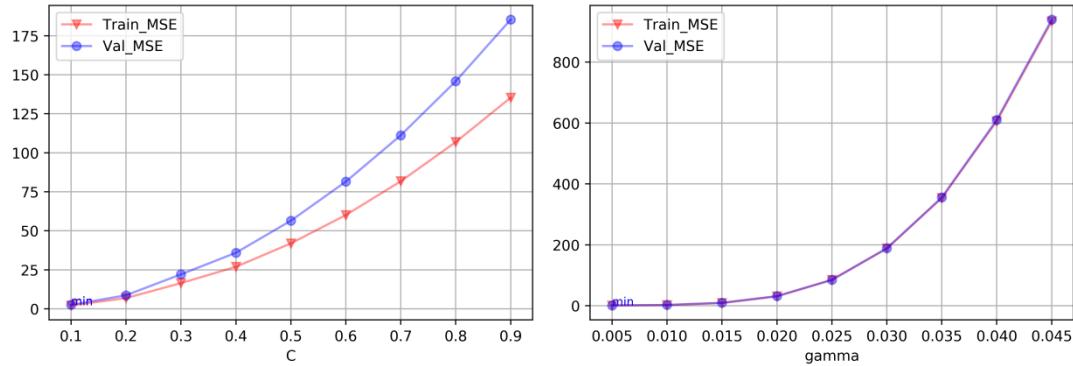


图 8 SVR-sigmoid 的超参数寻优过程

对于高斯核函数下的寻优，如下图。和多项式核函数一样，但其超参数只有惩罚参数 c 和核系数 γ 。同上述方法，我们发现高斯核函数的结果明显优于上述两者，其最优超参数惩罚参数 c 和核系数 γ 分别为 20 和 0.01。

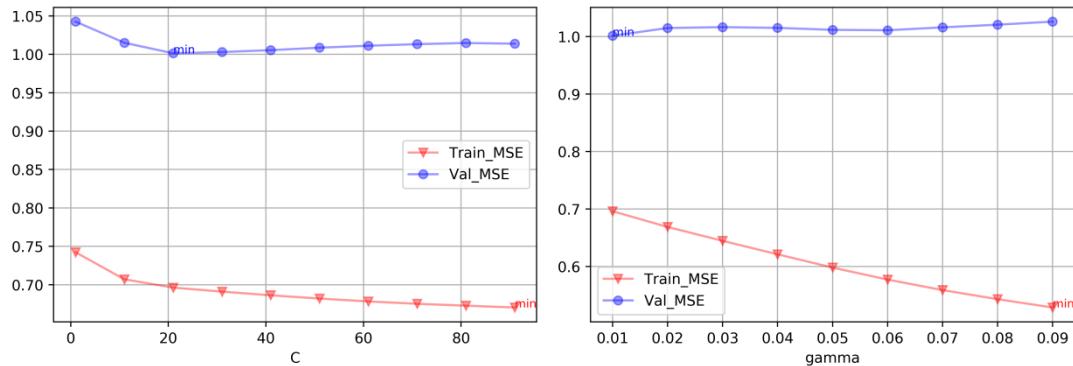


图 9 SVR-rbf 的超参数寻优过程

可以看出其中高斯核函数的结果优于多项式和 sigmoid 核函数，并且随着 γ 和 C 的变化，验证集的 mse 值先下降后降低，由此得到最优的超参数。高斯核（rbf）的表现较好，因此在后续分析时采用高斯核的 SVR 与其他模型对比。

4.4 模型预测结果

我们总共采用了 8 个模型，对每个模型在 12 个气象站的最优表现（注：在寻找模型最优超参数时以验证集 mse 作为指标）加以对比。每个模型在所有气象站下的表现见表格文件 test_data_of_each_station.xls。

定义每个模型最优表现：对每个模型的每个指标在 12 个气象站上剔除两个最差的之后，取平均值（也就是表格的最后一列）作为该模型最好结果。比较各个模型的指标，如下图所示。

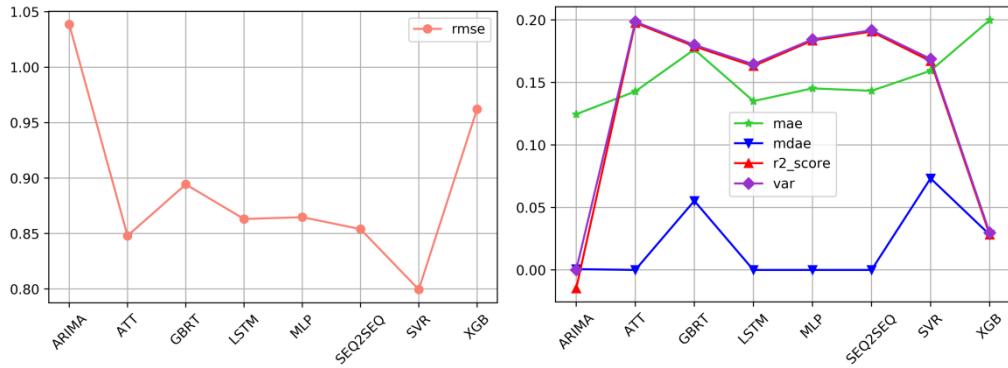


图 10 各个模型预测的平均评价指标对比

横坐标表示模型名称，其中 SVR 采用 rbf 核，纵坐标为五种指标，rmse 值相对其他指标较大，因此单独画出。

可以看出在 rmse 指标上，SVR (rbf 核) 模型表现最好，各个模型 mae、mdae 差别不大，在 r2 得分和可释方差得分上 ARIMA 和 XGB 表现很差，ARIMA 的 r2 得分甚至为负，XGB 的得分也很接近于零。

综合来看，单个模型中，神经网络和树、SVR (rbf 核) 模型比较稳定，在 5 个评价指标上也更好，而 ARIMA 和 XGBoost 模型不太稳定，从指标上来看效果也不好。

XGBoost 算法是 GBRT 算法考虑正则化的优化改进算法，但在本次实验中并未获得较大精度的提高，原因可能是本次降水量预测的数据非线性较强，而很难仅通过正则化来增强泛化性能。

整体上 seq2seq 网络的预测效果在多个指标上都是较好的，采用 attention 机制的 seq2seq 网络也更好的把握了在预测降水量上的特征。表中列举了数据集中 5 个气象站的结果，综合来看预测效果不随数据集有太大的变化。例如，在 5 个气象站的预测任务中，seq2seq 和 lstm 的 RMSE 分别比 ARIMA 模型下降 25% 和 24%，r2 得分达到了接近 0.3，seq2seq 和 lstm 的 RMSE 分别比 SVR 模型下降 9.5% 和 9.5%，r2 得分提升了接近 80%。seq2seq 和 lstm 的 RMSE 分别比 GBRT 模型下降 2.2% 和 2.1%，r2 得分上升 18% 和 14%。这主要是由于 ARIMA 和 SVR 等方法难以处理复杂的非平稳时间序列数据。FNN、GBRT 和 XGBoost 模型的

预测效果略低于 LSTM 是因为 FNN 只考虑了空间特征，而忽略了降水量数据是典型的时间序列数据。

4.5 模型集成后结果

在测试了单个模型的策略后，我们对其中神经网络构建的模型进行模型集成，来提高模型的泛化能力。ARIMA、XGBoost 模型不做模型集成的训练，是因为 ARIMA 的表现在各个气象站和指标上都不令人满意，而 XGBoost 算法本身就是利用了模型集成的思路构建，所以也并不需要继续做模型集成。

我们采用了两种方式构建模型集成方法，具体的方法在上文中已介绍，即 bagging 和 stacking。stacking 方法的次级学习器用到的是 GBRT。利用两种集成方法，在三个气象站（312、313、371）上，得到的 6 个集成结果如下。

下图为集成方法应用于 312 气象站上的结果。

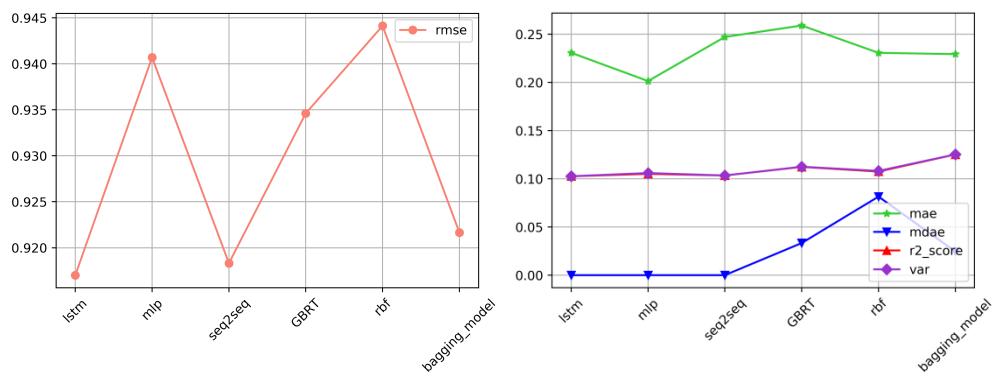


图 11 气象站 312 的各个模型的 bagging 结果

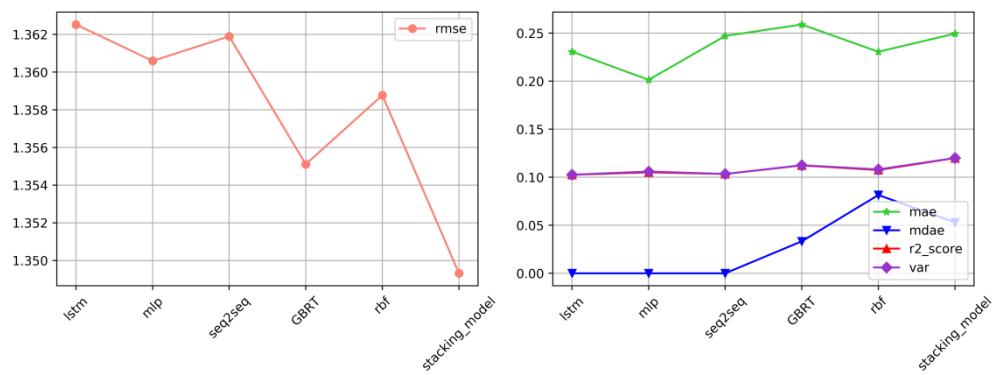


图 12 气象站 312 的各个模型的 stacking 结果

下图为集成方法应用于 313 气象站上的结果。

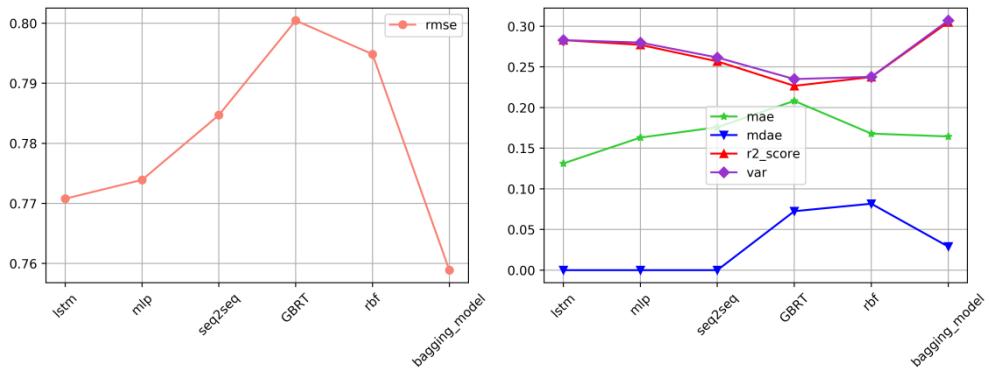


图 13 气象站 313 的各个模型的 bagging 结果

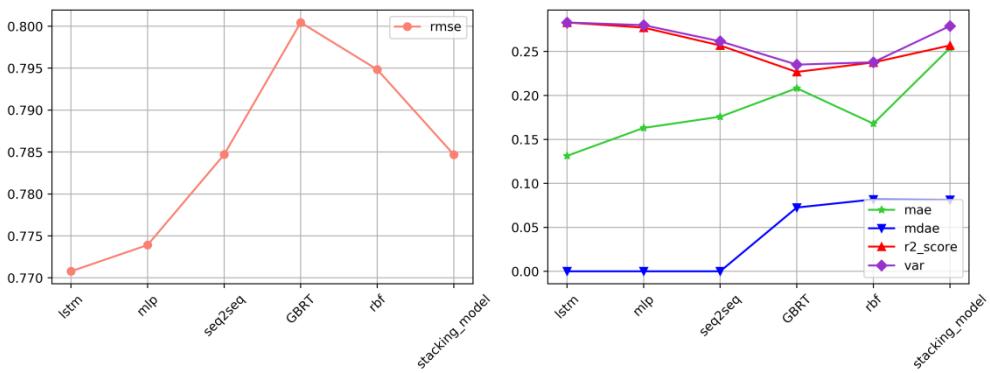


图 14 气象站 313 的各个模型的 stacking 结果

下图为集成方法应用于 371 气象站上的结果。

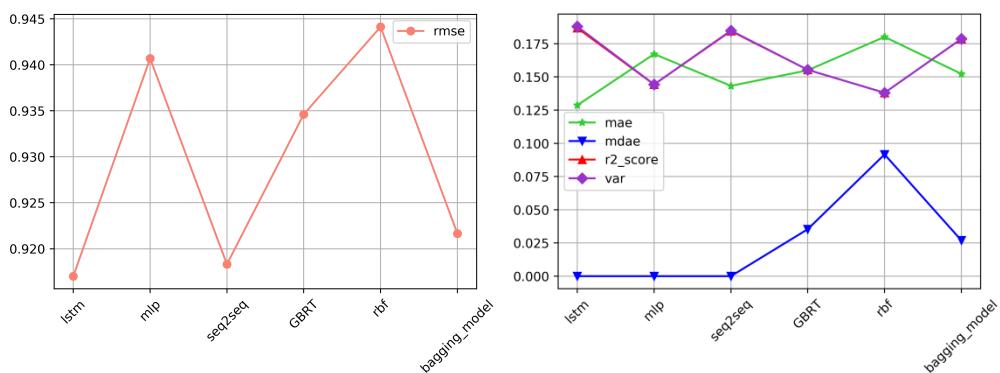


图 15 气象站 371 的各个模型的 bagging 结果

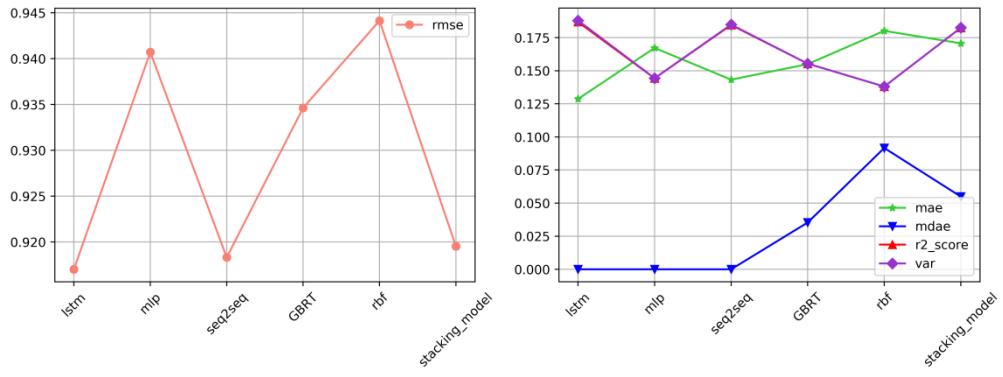


图 16 气象站 371 的各个模型的 stacking 结果

无论对于哪个气象站，对比于单个模型的结果，可以明显看出 ensemble 后的各项指标都不会比最差的单个模型的指标差，即至少保证平均性能，而且在有些参数上获得了一定的提升，超越了集成前的各个模型。

5. 实验结果分析

5.1 打乱数据集

在划分训练集和测试集时，打乱顺序再进行训练是否能得到更好的性能呢？对此，对于除 ARIMA 和 XGBoost 模型外的其他模型在每个气象站上我们都做了实验，每个模型在所有气象站上的表现详见表格文件 test_data_of_each_station_shuffle.xls。

首先，来看一下单个模型的最好表现对比，如下图所示。

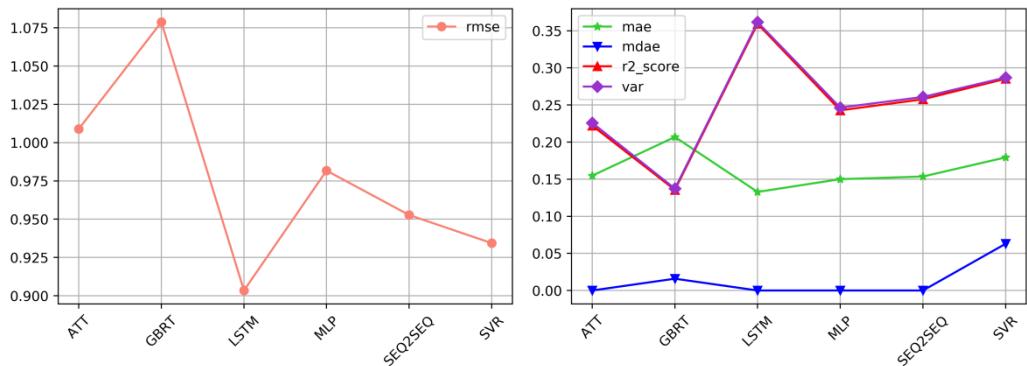


图 17 打乱数据集下的单独模型结果

可以看出，LSTM 表现最好，GBRT 表现最差。最好的 rmse 为 0.9 左右，r2 为 0.35 左右。平均来看 r2 得分在 0.25 左右，而相比于未打乱时的表现，如下图

所示，最好性能 ATT_Seq2Seq 的 r2 得分只达到 0.2。由此来看，模型在打乱的数据集上能够学习到更好的时空特征，高维度的拟合效果更优。注意到 seq2seq 网络在打乱的数据集上性能提升弱于 lstm，可能的原因是打乱之后特征的时序性并未改变。

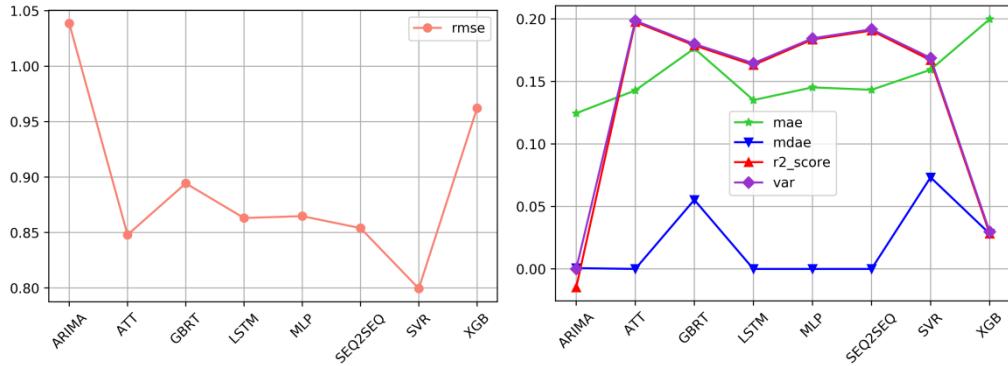


图 18 未打乱数据集下的单独模型结果

再来看同一模型在同一气象站下数据集打乱前后的可视化效果对比，如下左右分别对应未打乱和打乱后的 LSTM 模型、seq2seq 模型在 314 气象站的预测表现，可以看出打乱后的预测效果是比较好的。

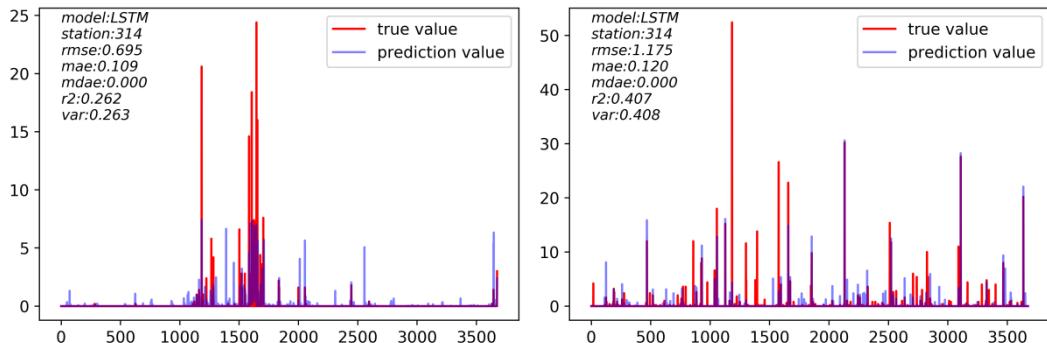


图 19 lstm 模型在未打乱和打乱的 314 气象站结果

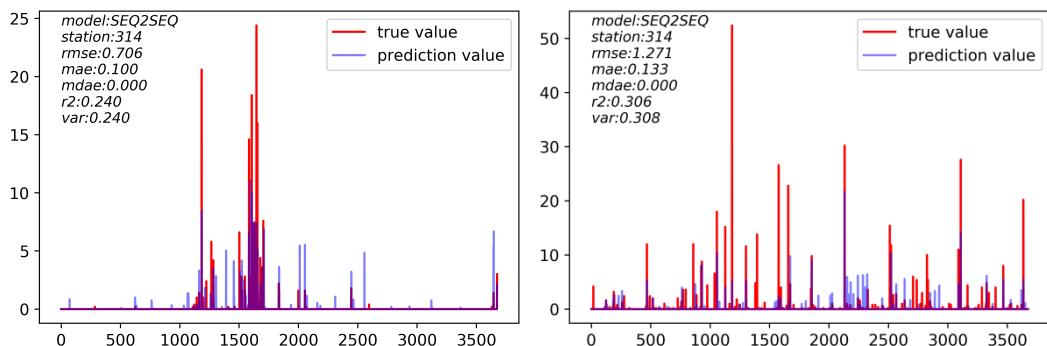


图 20 seq2seq 模型在未打乱和打乱的 314 气象站结果

但是这种不按照时空顺序进行预测的方式在训练的过程中加入了时间信息，更多的是体现“拟合”而不是“预测”降雨量信息。因此，在实际提交的最优模型中，我们提交的是未打乱情况下表现较为稳定的 seq2seq 的 5 个气象站的模型。但是这种方式也具有一定的实践价值，因为在实际情况下，当新的数据得到后可以继续用来训练来提升模型，所以在这样打乱的数据集上表现较好也说明我们的模型有效。

5.2 改变特征维度

在预处理数据集时，我们采取的是用前 3 个小时的 7 个维度的信息去做下一个小时的降雨量预测。那么，改变为用前 1 小时、2 小时的数据去做预测结果会怎么样呢？对此，我们也做了实验。每个模型在所有气象站上的表现详见表格 test_data_of_each_station_2_hours.xls 以及 test_data_of_each_station_1_hour.xls。下图是 2 小时预测的最优性能，每个模型的 r2 得分都在 0.15 之下（除了 MLP 略高于 0.15 之外）。相较于 3 小时的情况，r2 得分都在 0.15 以上至 0.20 之间，因此可以得出在单个模型的最优性能上用 2 小时的预测效果不如 3 小时的预测效果。

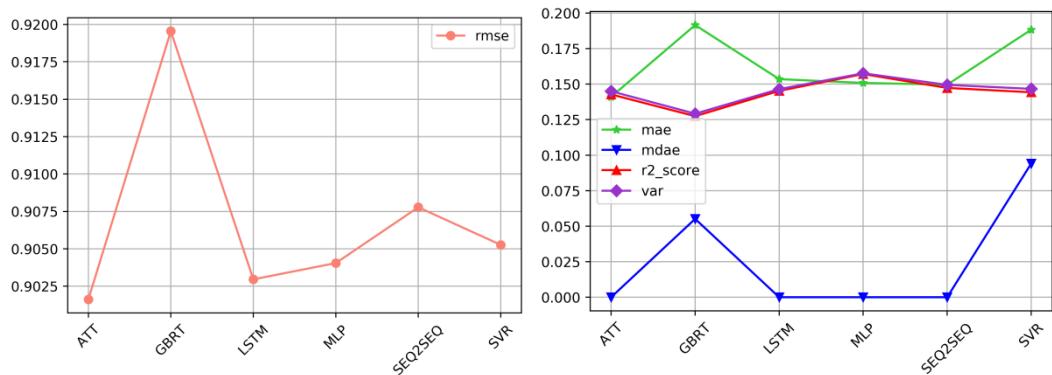


图 21 前 2 个小时数据的单独模型的结果

再进行 1 小时预测的实验时，由于时间关系，我们只做了三种模型在全部气象站上的最好表现，如下图所示。

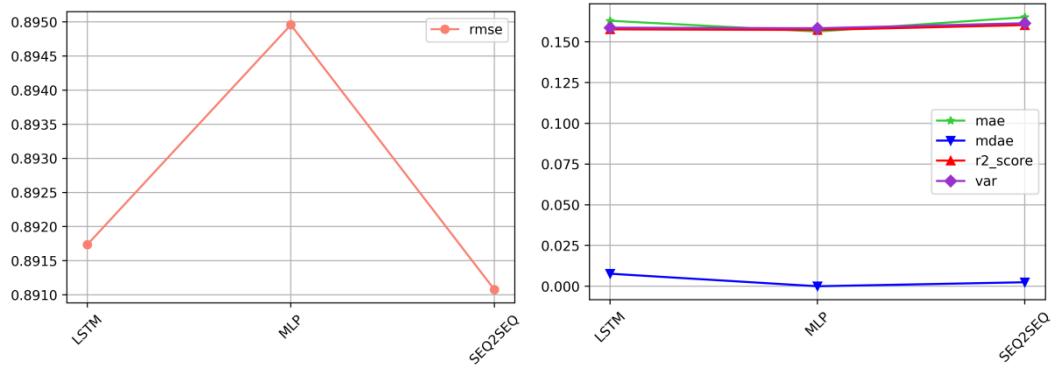


图 22 前 1 个小时数据的单独模型的结果

可以发现， r^2 得分在略高于 0.15 的位置，因此，只从 r^2 得分角度来看模型的最优效果排序为 3 小时>1 小时>2 小时，后续可以尝试做进一步的增加小时数的实验以及对结果作出一些合理的解释。

5.3 神经网络结构变化

上述的神经网络超参数寻优就已经涉及隐藏层个数以及维数，也就是模型的结构本身作为超参数，在探讨神经网络超参数寻优的地方已经可以看到较为明显的结论：网络层数和维度选择的大一些对降雨量的预测效果更好。相比于简单的 1 层 2 层而言，MLP 和 LSTM 都分别在隐层个数为 4 层和 3 层时达到最优。至少对于本数据集而言，神经网络的隐层个数为 3 层、4 层是比较合适的。

5.4 特征重要性分析

在本次实验中我们以 t 时刻的降水、太阳辐射强度、平均温度、平均露点温度、平均湿度、风向、阵风速度作为输入特征，对于特征的重要性分析，我们采用的方法是所构建的 seq2seq 网络，通过增加 Attention 机制来实现对序列的每个维度的重要性分析。我们输入前 3 个小时的气象特征，特征的维度就是序列的长度为 7。通过输出的前 5 个数据的热图判断输出与输入各维度的关系，这样可以得到每个小时内哪些气象参数更为重要，下图展示为三个气象站上前 5 个数据的注意力权重。

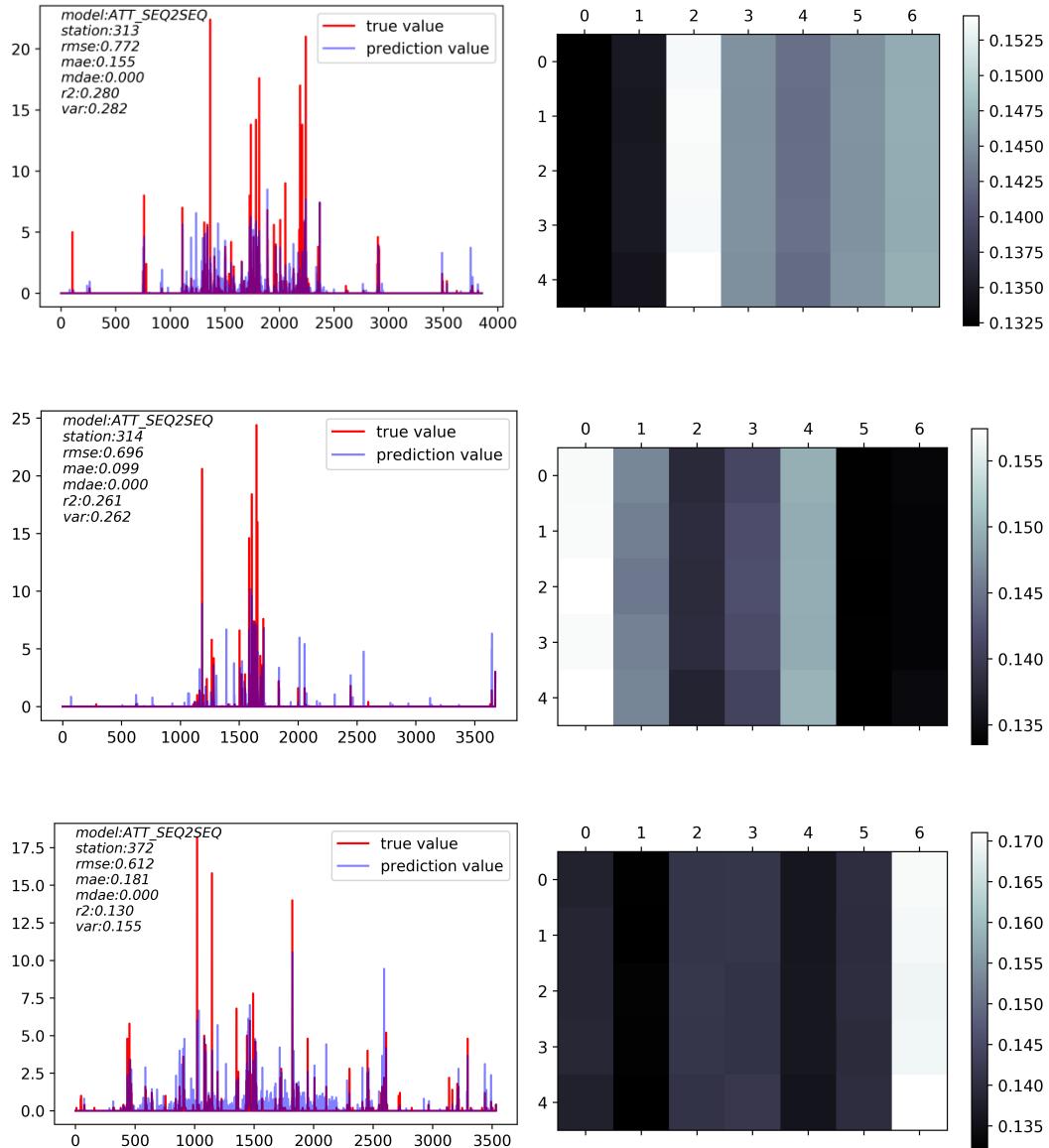


图 23 seq2seq 网络的注意力权重矩阵

可以看出，对于不同的气象站，7 个特征的重要程度是不完全相同的。通过观察可视化的数据，分析对于 314 气象站，其有降水的时间较少，模型对 t 时刻的降水量和平均湿度关注更高；对于 313 气象站，其有降水的时间中等，模型对 t 时刻的平均温度、阵风速度关注较高；而对于 372 气象站，其有降水的时间较多，模型对 t 时刻的阵风速度主要关注阵风速度。这和常理是符合的，当此地的降水时刻较少时，一般降水是密集的，前一时刻的降水有无对下一时刻有很大的影响。但如果此地属于降水时刻多的地方，一般是有对流天气较多，那降水的有无很大程度上和前一时刻的阵风有关，在皮尔逊系数中也发现两者的关系比较密

切。但对于其中的机理，后续可以增加选取不同维度的输入信息去做降雨量预测的实验。

5.5 模型及可视化分析

从单个模型在所有气象站上的综合表现来看，seq2seq 和带 attention 机制的 seq2seq 模型表现最好、最稳定。选取 seq2seq 模型在几个气象站上的可视化结果加以分析。如下图，为 seq2seq 在四个气象站上的预测结果。

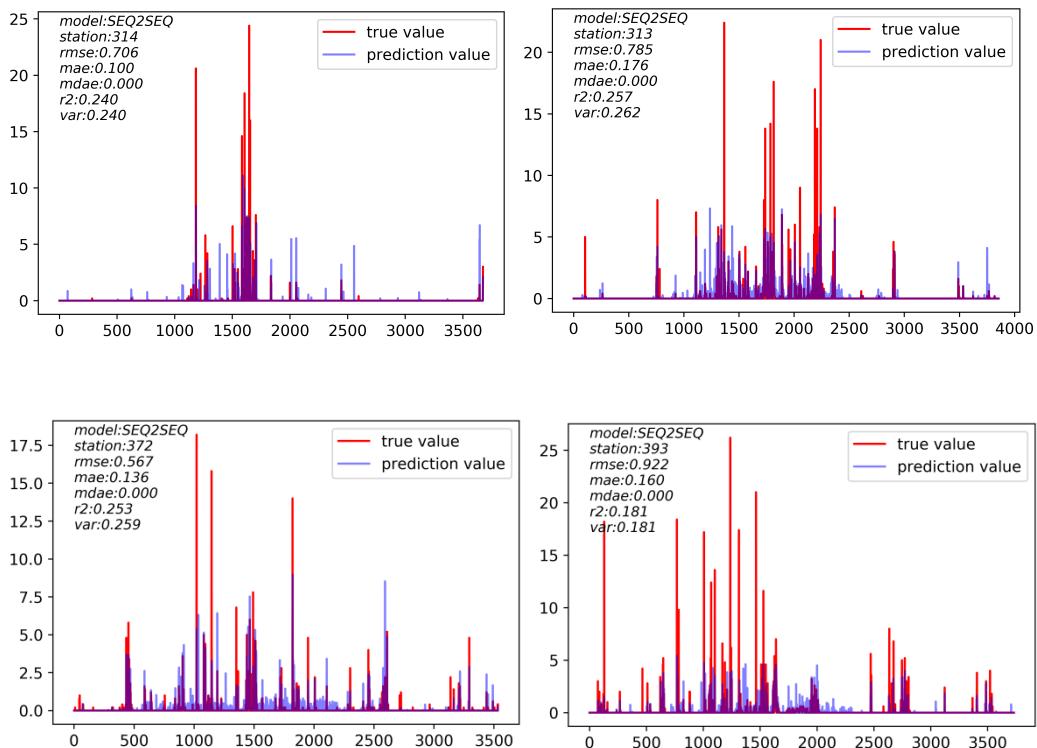


图 24 seq2seq 模型的预测结果

可以看出 seq2seq 模型基本能够预测到降水量的变化，能够在集中有雨的时候判断出降雨，但其在面对偶然的大幅度波动时，模型很难预测精准。虽然对于一些较大的降水量没有预测的非常准确，但其在没有雨的时候也很少判断出有降雨，能够在一定程度上对小范围的波动预测较好。对于 seq2seq 模型，我们将其复杂度设计的较低，得到的泛化性能较好。

6. 结论

本文我们针对降水量预测问题，首先对降水量时序预测问题进行建模，利用皮尔逊系数得到对降水量相关系数较强的气象参数。然后利用 ARIMA、SVR、

GBRT、XGBoost、FNN、LSTM、seq2seq、attention 机制的 seq2seq 模型对问题进行求解，探究了不同超参数下的寻优过程，实现了对下一个小时的降水预测，并在多个气象站得到了验证，并对比了不同模型的预测效果。经过测试，单个模型中最为稳定的是 seq2seq 模型，经过模型集成中的 bagging 和 stacking 方法使得模型的预测性能进一步提高。此外我们还研究了不同输入序列长度对预测效果的影响，发现输入前三个小时的输入优于两个小时和一个小时，从 r2 得分上看提升接近 20%；在使用非时序的打乱的数据集训练和测试后，模型效果也进一步提升了 30%左右；使用 attention 机制的 seq2seq 模型来可视化不同地区所训练模型的特征重要性，通过分析其中的区别为我们后续继续改善模型打下基础。

致谢

这项工作得到了清华大学模式识别课程的张长水老师和各位助教的大力支持。由于我校在新冠病毒流行期间入校审核严格，我们感谢清华大学导航、制导与控制实验室的师兄师姐协调的研究资源。部分计算是在清华大学导航、制导与控制实验室服务器（“彭于晏”）系统上完成的。

小组成员贡献

王圣杰负责数据的预处理和基本模型的搭建，并主要负责对于单独模型的调优实验，及作业报告中实验开始以前的部分的撰写，并负责撰写 readme.md。董宇光主要负责模型的集成和数据后处理任务，以及相应的数据分析部分的撰写，也参与部分模型的调优工作。

代码链接

代码主要由组内成员独立编写，但对于 ARIMA、SVR、GBRT 和 XGB 模型的搭建是参考 sklearn 和 statsmodels 模块辅助编写，使用 Pytorch 库对各个神经网络模型进行搭建。具体见附带的 README.md 文件。

参考文献

- [1] 沈皓俊, 罗勇, 赵宗慈, 王汉杰. 基于 LSTM 网络的中国夏季降水预测研究[J/OL]. 气候变化研究进展: 1-18[2020-05-22]. <http://kns.cnki.net/kcms/detail/11.5368.P.20200306.2028.003.html>.
- [2] Chinchorkar SS, Patel GR, Sayyad FG (2012) Development of monsoon model for long range forecast rainfall explored for Anand (Gujarat-India). *Int J Water Resour Environ Eng* 4(11):322–326
- [3] Schepen A, Wang Q J, Robertson D E. Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall [J]. *Journal of Geophysical Research: Atmospheres*, 2012, 117(D20)
- [4] Yevjevich, VujicaM. Stochastic processes in hydrology[M]. Water Resources Publications, 1972.
- [5] Matalas NC, Wallis JR (1971) Statistical properties of multivariate fractional noise process. *Water Resour Res* 7:1460–1468
- [6] Carlson RF, MacCormick AJA, Watts DG (1970) Application of linear models to four annual streamflowseries. *Water Resour Res* 6(4):1070–1078
- [7] Valencia DR, Schaake JC Jr (1973) Disaggregation processes in stochastic hydrology. *Water Resour Res* 9(3):580–585
- [8] Moustris KP, Ioanna K, Larissi (2011) Precipitation forecast using artificial neural networks in specific regions of Greece. *Water Resour Manag* 25:1979–1993
- [9] Ramana R V, Krishna B, Kumar S R, et al. Monthly rainfall prediction using wavelet neural network analysis[J]. *Water resources management*, 2013, 27(10): 3697-3711.
- [10] Bartoletti N, Casagli F, Marsili-Libelli S, Nardi A, Palandri L (2018) Data-driven rainfall/runoff modelling basedon a neuro-fuzzy inference system. *Environ Model Softw* 106:35–47
- [11] Cramer S, Kampouridis M, Freitas AA, Alexandridis AK (2017) An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Syst Appl* 85:169–181
- [12] Lin GF, Jhong BC (2015) A real-time forecasting model for the spatial distribution of typhoon rainfall. *J Hydrol* 521:302–313
- [13] Pham, Q. B. , Abba, S. I. , Usman, A. G. , Thi, N. , & Tri, D. Q. . (2019). Potential of hybrid data-intelligence algorithms for multi-station modelling of rainfall. *Water Resources Management***, 33(15), 5067–5087.
- [14] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine- grained air quality based on big data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2267–2276. ACM, 2015.
- [15] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, Yu Zheng, GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction. In International Joint Conference on Artificial Intelligence (IJCAI), 2018.
- [16] Zhao L , Song Y , Zhang C , et al. T-GCN: A Temporal Graph ConvolutionalNetwork for Traffic Prediction[J]. 2018.