降雨量预测

张芙作 2019211220
zhangfz19@mails.tsinghua.edu.cn

摘要

关键字：

1 简介

降雨量任务属于时空序列类预测问题，由于该问题的研究具有实际的应用价值和意义，所以一直以来备受学界和工业界的关注。气象参数之间往往存在这非常复杂的相互关系，而且降雨量本身存在不确定性和随机性，因此降雨量预测是一个非常有难度的任务。降雨量预测中，根据研究思路的区分主要有数值天气预报模型和遥感观测两种方法[1]，后者主要是基于专业领域的气象观测而获得的，在这里不作讨论。

基于天气数据对降雨量进行预测的方法早期有基于统计模型的线性模型、多元线性回归等，之后随着机器学习算法的发展和广泛应用，有研究者用人工神经网络进行短期的降雨量预测。由于用单个模型进行预测结果往往不尽人意，有学者采用混合模型的方法来实现降雨量预测，比如将K近邻，人工神经网络模型的结果通过加权平均来进行预测，从而避免单个模型的不稳定性。文献[3]分别用线性回归模型和神经网络来对局部以及全局气象数据进行建模，并分别进行预测，并通过动态的组合方法将两种预测器预测的结果进行组合，基于这样的组合模型对空气质量中的PM2.5进行预测，并给予这样的模型实现了在线的预测工具。文献[4]结合了小波变换技术和神经网络模型来解决降雨量预测问题。不仅仅是传统的机器学习算法，近年来由于深度学习在时间序列问题上有了非常成功的应用开始有研究者将深度学习模型引入到时间序列的降雨量预测问题上，以实现端到端的建模和预测。文献[5]中提出了一种用于时空序列数据处理的深度网络预测模型，引入了多层卷积神经网络来提取时间上的变化趋势和空间上的距离相关性，并将时空特征融合进一步预测。文献[2]中采用了双层LSTM网络结构来实现基于时空序列数据的空气质量预测和降雨量预测，为了提高模型的拟合能力，同时引入了注意力机制对不同的特征赋予不同的权重，从而自动提取其他检测变量与目标变量之间的动态相关性。与其他的基于统计模型或者智能算法的方法相比，采用如LSTM这种循环神经网络的方法可以方便的对时间序列进行处理，无需过多的数据预处理，可以完整保留数据在时间上的变化特性和变量在时间上的相关性。所以越来越多的学者开始尝试用循环神经网络来处理时空序列问题。

在本次项目中，我们将降雨量预测问题定义为通过前六时刻的气象数据对当地下一时刻的降雨量类别进行预测的分类问题，我们分别采用了线性分类器、基于决策树的分类算法以及基于神经网

络的模型来解决降雨量预测问题，并通过对不同模型的组合来进一步得到拟合能力更高的模型。我们对五个在同一个城市的不同气象站分别用模型对当地的气象特点进行建模，分别对当地的降雨量进行预测，并对采用了不同模型的分类器结果进行了比较和分析。

2 任务定义

本文将降水量预测问题定义为分类问题，按照表1中所示的标准根据降水量的数值分成无雨，小雨，中雨和大到暴雨四类:

表 1: 降雨量分类标准

降雨量(毫米/小时)	0	(0,1]	(1,4]	(4,∞)
类别	无雨	小雨	中雨	大到暴雨
标号	0	1	2	3

同时，本文将问题定义为短时预测问题，基于前六小时的特征数据预测下一小时的降水量，对 5 个邻近气象站分别单独建模。

获得的原始数据为每个气象站按照时间顺序每个时刻（一小时一个采用）的气象特征监测值，从第一个有记录的时刻开始进行窗口滑动，窗口滑动的步伐为1，当前时刻到之后的连续第6个时刻的数据样本作为一个样本，第7时刻的降雨量类别作为该样本的目标值。这样将每一个样本都是一个二维的$T \times P$矩阵，其中T为时间周期，在这里为6，P 为每一个时刻所选用的特征个数。所以当共有N个样本时，降雨量预测任务的输入为N个样本的特征，即为$N \times T \times P$的数据集，希望降雨量预测模型能够根据每一个样本给出该样本所示的前六个时刻的气象数据给出下一时刻的降雨量类别预测，这样模型对所有N个输入样本的输出为一个长度为N的类别向量。

3 数据整理及特征处理

3.1 数据整理

数据来源及内容：

数据清洗，缺失值处理方法：

数据分布的可视化：

3.2 特征处理

数据归一化处理：

特征间的组合：

使用的特征维度和每一维的含义：

4 模型设计

这里至少四种不同的单独模型

4.1 线性模型

4.2 XXX

4.3 Xgboost

XGBoost是Boosting算法的由多个弱决策分类树组成的一个强分类器。类似于基本的CART树一样，通过不断地特征分裂来完成一棵决策树的构造，为了对给定的损失函数进行最小化，采用贪心算法对每次节点分裂时的特征进行选择。与传统的CART树采用不纯度等损失函数构造决策树不同，XGBoost中决策树的构造可以根据需要设置目标函数，并且一般来讲，目标函数由两部分组成。第一部分衡量预测类别和真实类别之间的差距，而另一部分则是正则化项，用于控制模型的复杂度，防止模型出现过拟合，提高了最终分类器的泛化能力。 XGBoost基于Boosting思想采用串行方法产生多个决策树，每一次添加新的树，都是去拟合上一次预测的残差，这样通过多次迭代产生一系列决策树组成一个大的分类器。

XGBoost算法是机器学习界的一大利器，被广泛应用于大数据竞赛和工程，其作为一种特殊的集成树算法，拥有许多出色的优点：

(1) 增加了很多防止过拟合的策略，比如目标函数中增加正则化项，对数据进行随机采样等；

(2) 在训练新的决策树时利用了损失函数的二阶导数，加快了优化速度

(3) 虽然基于Boosting思想树与树之间是串行关系，但是单个树的节点分裂过程可以实现并行化，用多线程来选择最佳分裂点，非常有效的加快了训练速度。

当然，作为一种基于决策树的算法，与众多树算法相同，XGBoost的决策树训练也采用了贪心算法，相对来说增加了训练过程的时间，另外由于XGBoost算法参数较多，在实际应用中需要较多的精力用来调参。

4.4 LSTM

LSTM（长短时记忆网络）是一种特殊的循环神经网络，其被广泛用于时间序列问题，特别是在自然语言处理问题上有非常突出的表现。由于LSTM网络的输入具有明确的时间先后关系，所以适合处理本项目中具有明显时间轴信息的时间序列数据。为了根据前六小时的天气状况预测当地下一时刻的降雨量，用多层的LSTM网络和全连接网络对该问题建模，图1给出了两层串行LSTM网络结构和最后输出连接全连接层的网络结构示意图。第一层LSTM单元按时间顺序依次接收输入数据，第一层各个单元的输出分别作为对应第二层网络单元的输入，最后一层LSTM的输出被简单组合为一个大的固定尺寸的向量，并构建全连接层用于实现分类任务。

训练集按照样本维度进行了归一化(零均值化和方差归一化)，并按照训练集上的参数对验证集和测试集进行同样的归一化操作，以使数据尺度相同。

4.5 集成算法

这里至少两种不同的集成算法，一下列举的是目前想到的ensemble方法

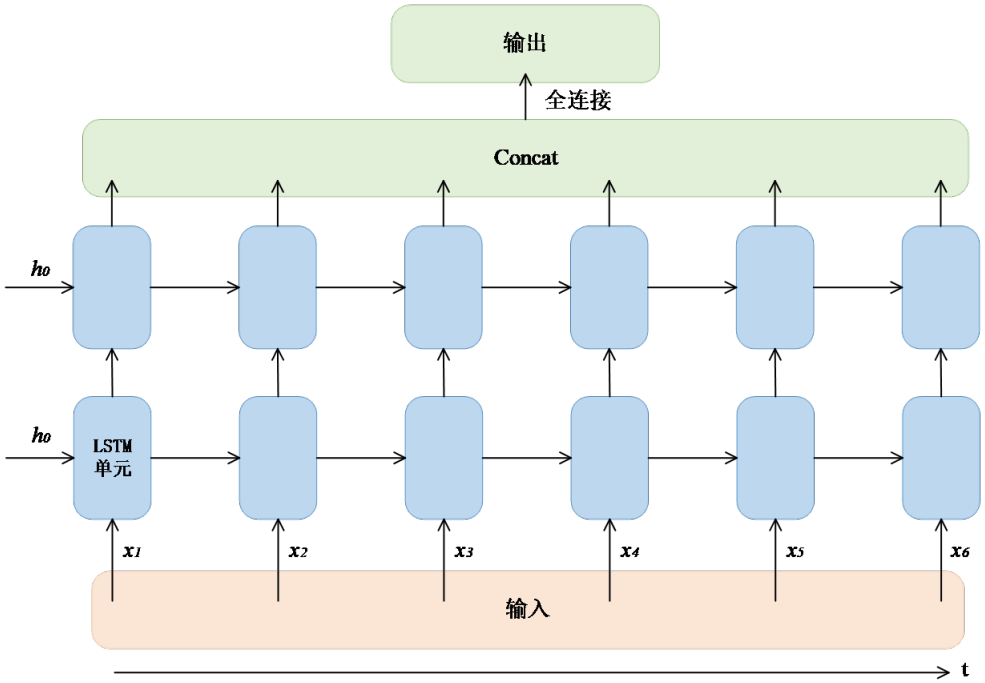

图 1: LSTM网络结构图

4.5.1 Bagging

用Bagging方法训练多个单独的分类器，并将分类器进行组合，以实现一个更强性能的大分类器。

4.5.2 Adaboost

5 实验设计及结果

5.1 训练集、验证、测试集的划分

由于天气特性与地理位置之间具有强的相关性，所以处于不同地理位置上的气象站的天气指标具有不同的分布特点，为了能够对各个气象站所在位置的降雨量进行更加精准的建模和预测，对每个气象站的降雨量预测任务单独建模。将每个气象站的天气数据按照时间顺序以7:2:1的比例划分为训练集、验证集和测试集。

为了避免训练集和验证、测试集中的数据出现交叉，影响模型性能的判断，先将清洗过后的数据拆分为不同集合，然后在每个集合中分别进行窗口滑动，从第一个有记录的时刻开始，窗口滑动的步伐为1，当前时刻到之后的连续第6个时刻的数据样本作为一个样本，第7时刻的降雨量类别作为该样本的目标值。

5.2 评价指标的定义

本次任务是一种多分类任务，所以选择一下四种评价指标对模型预测能力进行判别：

1. 准确率(accuracy)：即分类正确的样本数站总样本数的比例，这是最为常用的分类问题评价标准，但是对于本问题中类别不平衡的数据集该评价指标不能有效说明分类器的预测性能，因为数据中降雨量类别为0的数据占90%左右，即使模型给出全0的结果也可以得到比较高的accuracy。

2. 精确率(precision)：二分类问题中，被关注的类的精确率计算方式为$p = \frac{TP}{TP+FP}$，本问题为一个多分类问题，计算各类的加权平均精确率作为多类别分类器在测试集上的最终精确率。

3. 召回率(recall)：与精确率的计算方式类似，计算各类的召回率$r = \frac{TP}{TP+FN}$，将各类的加权平均召回率作为模型在测试集上的最终召回率。

4. f1-score：精确率和召回率是一对相互矛盾的量，有着相反的变化趋势，所以为了能够更好的评价分类器的性能，使用f1-score= $\frac{2\times p\times r}{p+r}$评价分类器的综合性能，多分类器的f1-score 为各类该指标的加权平均。

5.3 单独模型的最好结果对比

XGBoost和lstm均为在清洗过后的数据上进行实验，并且对每个气象站进行建模时，去掉了整个数据中为常量，或者冗余的特征变量，共保留了21个特征变量。

LSTM网络模型在迭代20次之后，损失函数就呈现非常缓慢的下降，继续增加迭代次数对最终模型的性能影响不大，所以共迭代30次。

表 2: 单独模型的最好结果对比

模型	accuracy	precision	recall	f1-score
LR	.93|.93 |.89|.96|.95	.90|.90|.83|.93|.93	.93|.93|.89|.96|.95	.91|.91|.84|.94|.93
SVM	.93|.93|.89|.96|.95	.91|.90|.84|.93|.92	.93|.93|.89|.96|.95	.91|.91|.85|.94|.92
KNN	.93|.92|.88|.96|.95	.90| .89|.83|.93|.91	.93|.93|.88|.96|.94	.91|.90|.85|.94|.92
DT	.93|.93|.90|.96|.95	.91|**.92**|.87|**.95**|.93	.93|.93|.90|.96|.95	.92|.92|.88|.95|.94
RF	.93|.93|.90|.96|.94	.91|.91|.87|.94|.93	.93|.93|.90|.95|.94	.92|.92|.88|.95 |.94
GBDT	.94|.93|.90|.96|.95	**.92**|**.92**|**.88**|**.95**|**.94**	**.94**|.93|.90 |.96|.95	**.93**|.92|**.89**|.95|.94
MLP	.94|.93|.90|.96|.95	**.92**|.91|.87|**.95**|**.94**	**.94**|.93|.90|.96|.95	**.93**|.92|.88|.95|.94
XGBoost	.93 |.93|.90|.95|.95	.92|.92|.87|.93|.94	.93|.93|.90|.95|.95	.92|.92|.88|.94|.94
LSTM	.93|.93|.88|.95|.95	.86|.86|.78|.89|.89	.93|.93|.88|.95|.95	.90|.90|.83|.92|.92

5.4 集成模型的最好结果对比

用Bagging方法对XGBoost模型进行集成，每一次从训练集中有放回的随机抽取和训练集一样大小的子训练集，用该子训练集训练一个相对简单的XGBoost分类器，并重复以上过程n次，最终得到n个相对较弱的XGBoost分类器，并基于投票的思想将n个弱分类器组成一个大的分类器，用验证集对该分类器进行验证，选择最好的分类器。

可以看到用多个由随机子训练集训练得到的XGBoost组合分类器在该问题上的性能和单独的最优XGBoost分类器并无显著差别。同样的方法，在随机子训练集上训练得到了多个LSTM分类器，并组成一个大的分类器，该分类器在测试集上的分类性能与单独的最优LSTM相比几乎没有提高，仍

表 3: 集成模型的最好结果对比

模型	accuracy	precision	recall	f1-score
Bagging-XGBoost	.94|.93|.90|.95|.95	.92|.92|.87|.93|.94	.94|.93|.90|.95|.95	.92|.92|.88|.94|.94

然出现了非常严重的类别不平衡带来的问题，组合分类器对无雨类别的样本能够给出高精度的预测结果，但是对于其他类别的样本几乎没有预测能力，精确率和召回率都为0，故此处不再给出单独用多个LSTM组合而成的大分类器具体表现结果。

5.5 每种模型在不同超参数下的表现

最初用XGBoost模型对每一个气象站的数据分别调参，希望对每一个气象站分别得到一个最优超参数。在实验过程中，发现XGBoost模型性能在该问题上对超参数的选择不敏感，每一个气象站的训练学习率从0.01到0.2调整，决策树最大深度从5到10变化，决策树个数从10到500变化，分别用单独的气象站数据训练得到的模型在验证集上的实验结果几乎没有差别，所以可以在流程上进行简化，对所有的气象站数据分别用同一组超参数进行训练，并为了节省模型的训练时间和测试时间，将模型超参数的设置在保证模型性能的基础上使最终模型尽可能的小。

LSTM模型的超参数有LSTM层数、隐层变量数以及训练过程的学习率上，通过多次实验进行选择，发现LSTM层数的增多对最终模型的拟合能力的提高没有明显的贡献，所以设置LSTM的层数为2，全连接层的增加也无法提高模型的拟合能力，所以在最后一层的 LSTM输出以上设置一个全连接层，LSTM单元的隐变量个数均为10。最初选择用带有动量项的随机梯度下降方法训练，无论怎样设置学习率和动量因子，模型都非常缓慢的收敛，最终选择了可以在训练过程中通过计算梯度的一阶矩估计和二阶矩估计而不断自适应更新学习率的Adam算法。调整初始的学习率发现，该训练过程对初始学习率的设置并不敏感，使学习率从0.01到0.2变化，模型训练速度基本相同，都能在30次迭代后基本达到收敛状态。

5.6 其他实验设计及结果

在实验中发现LSTM模型对这种类别不平衡问题表现的不好，特别是对于数据较少的类别，所以希望能够通过一些减少类别不平衡问题的操作，提高LSTM模型对该问题的拟合能力和预测能力。EasyEnsemble算法是一种用于处理类别不平衡问题的欠采样算法，在本问题上该算法的实现步骤如下：

(1) 从无雨的类别数据中又放回的随机采样n次，每次选取与其他类别数据数目相近的样本个数，最后得到n个用子样本集合$\{S_1, S_2, \ldots, S_n\}$。

(2) 分别将以上n个子样本集合与其他类别的数据样本合并为n个子训练集，并用每个子训练集训练出一个LSTM模型，如此可得到n个模型。

(3) 将这n个模型组合成一个集成学习系统，最终用这n个模型的结果进行投票。

在实验中，将每次对无雨的类别数据采样数据为其他类别数据总和的0.4倍，一共生成100个子训练集，从而训练得到100个子分类器。不过得到的训练器整体性能与单独的LSTM模型相比并没有

提升，反而下降。以在气象站734上用EasyEnsemble方法训练得到的多LSTM模型组成的大分类器为例，该分类器在测试集上的实验结果如表4中所示。

表 4: EasyEnsemble方法在气象站373测试集上的实验结果

类别	precision	recall	f1-score	样本量
无雨(0)	0.97	0.50	0.66	7046
小雨(1)	0.12	0.89	0.22	595
中雨(2)	0.00	0.00	0.00	228
大雨到暴雨(3)	0.00	0.00	0.00	95
加权平均	0.87	0.51	0.60	7964

该分类器在测试集上的预测准确率为0.51，相比于表3中单独LSTM模型的预测性能，该分类器的性能变弱了。precision，recall和f1-score的加权平均值都比单独LSTM模型的测试结果小，但是分开看各个类别的预测结果，其中表4中显示该分类器对小雨类别的数据预测 recall为0.89，这个是明显强于单独LSTM模型的，前面已经介绍过，用所有训练数据得到的单独LSTM模型除了对无雨类别数据能给出很高的预测结果，其他类别的数据预测能力几乎为0，即其他类别的评价值均为0。而用EasyEnsemble 对无雨类别数据欠采样得到的大分类器能够对小雨类别的数据有一定的预测能力，这是该方法于单独LSTM方法提高的点，这说明用EasyEnsemble方法得到的一系列子分类器组成的大分类器在一定程度上能够缓解类别不平衡问题，不过只是在一定程度上。如表4中中雨和大雨到暴雨类别的数据分类器预测的评价指标仍均为0，分析原因可能是因为小雨类别的数据要显著高于中雨和大雨到暴雨这两类，在对无雨类别数据欠采样后，子训练集中小雨和无雨两类数据占比较高，而另外两种就占比很少，所以最终得到的分类器对后面两类不能给出正确的预测结果，所以这样看来，采用EasyEnsemble算法只在一定程度上解决了部分类别不平衡的问题，由于四类数据之间的样本量差距都比较大，所以无法用这种对多类数据欠采样的方法来从根本上解决这个问题，而且由于欠采样后子训练集中的样本量较少，难以用来训练得到较为高性能的LSTM模型，进而得到的大分类器性能也较差，故该实验没有在ensemble部分进行报告。

6 结果分析与讨论

6.1 数据和模型中每一部分的贡献

在本项目中，我们分别对每一个气象站的降雨与其他气象特征单独建模，用前六时刻的气象特征预测下一时刻的降雨可能性与降雨量大小。数据特征的选取及各特征与目标变量之间的相关性分析可见下一节中关于特征的具体分析。

下面分别讨论各个模型在解决降雨量预测问题上的作用。

1. XGBoost模型是一种基于Adaboost思想的决策树森林，森林中的树按照串行关系依次用训练集训练得到，每个树的优化目标都与之前已生成的决策树有关。最后得到的XGBoost 模型包含多个子分类器，在对输入进行预测时，每个子分类器都给出各自的预测结果，最后所有子分类器进行投票，最终给出XGBoost模型的分类结果。

2. LSTM模型是一种循环神经网络模型，与其他的算法不同，使用该模型进行降雨量预测无需

将$T \times P$格式的二维矩阵展开为一维向量，按照时间顺序将T个时刻的特征向量依次输入到LSTM网络中。用于降雨量预测的LSTM模型由两部分组成，输入的特征向量先经过两层LSTM网络进行特征提取，第二层LSTM网络的输入为一个长度固定的向量，该向量再输入一个全连接网络中，用于给出分类结果。在整个LSTM模型中，两层LSTM网络作为一个特征提取的过程，并将特征用一个固定长度的向量来表示，后面的全连接网络作为结果预测部分，用提取到的特征对最后分类结果进行预测。

6.2 特征的重要性分析

6.3 错误分析

（模型在哪些数据下预测准确率高，哪些数据下预测准确率低）。

如前面所介绍过的，由于实验中所采用的地区降雨都为小概率时间，所以大部分时间都为无雨状态。即使下雨，大部分情况为小雨，中雨出现次数更少，而大雨的概率就极低。使得该问题中不同类别的数据样本量差别极大，属于严重类别不平衡的情况。无雨的类别占比能达到90%以上，所以用该数据训练得到的预测模型，对于无雨的情况能够给精确度和召回率都非常高的预测结果，而对于其他三种类型的数据预测能力就会弱很多。为了具体说明这种情况的普遍性，表5示出了气象站375上的最优XGBoost模型在其测试集上的实验结果。在该数据集上，无雨的样本占据94.6%，其他三类数据共占据5.4%。无雨类别的precision和recall值分别为0.96，0.99，f1-score为0.98，都非常高，说明该该模型对无雨类别的数据有很高的拟合能力，这与训练集中绝大部分样本也为无雨类别有关。而测试集中小雨、中雨和大雨类别的f1-score值都在0.5以下，这说明该模型对于这三类样本的拟合能力比较弱，特别是对雨样本量不足2%的中雨和大雨两类，预测能力非常弱。

类别不均衡带来的分类器在不同类别上表现具有巨大差异的问题在LSTM模型上更为明显，除了在气象站373上的模型，其他气象站的LSTM模型在测试集上中无雨类别上的f1-score均可达到0.95及以上（气象站373为0.94），而在其他类别上的precision和recall几乎都为零，即LSTM模型完全无法对这种在训练集中占比极低的样本进行拟合。

表 5: 最优XGBoost模型在气象站375上的测试结果

类别	precision	recall	f1-score	样本量
无雨(0)	0.96	0.99	0.98	7992
小雨(1)	0.46	0.26	0.34	330
中雨(2)	0.27	0.07	0.12	95
大雨到暴雨(3)	0.75	0.09	0.15	35

6.4 案例分析

（在具体的案例上，不同模型表现的区别在哪里）。

模型和结果可视化分析。

其他方式的实验结果分析。（如果有的话）

7 小组成员贡献

8 结论

参考文献

[1] 周泽世. 基于BP神经网络的降雨量预测研究[D].湖南农业大学, 2015.

[2] 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.

[3] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting finegrained 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.

[4] 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.

[5] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. Dnn-based prediction model for spatio-temporal data. ACM SIGSPATIAL 2016, October 2016.