

基于决策树模型和神经网络模型的降雨量问题研究

李沐阳

易领程

钟绍恒

2024 年 5 月 7 日

1 摘要

本文针对局部地区的降雨量与其它气象相关的因素进行了数学模型的建立和求解的通用算法的设计,运用线性回归,障碍计数,时间序列和神经网络等模型分别进行了尝试,最终发现决策树取得了较好的结果.

根据气象预测的要求,定义正确为与真实数据相差的绝对值小于 10mm. 我们首先针对与降水量关联度最大的水平面气压¹和气温等条件,将数据按照 30 天为一组进行平均处理整理后,根据气象学上的基本原理,运用线性回归模型拟合真实数据,完成了该简单模型的求解,取得了正确率 53% 和绝对平均误差 11mm 的良好的预测效果,在这里同时使用了最小二乘法进行拟合,均能够取得很好的效果. 在此时还尝试了障碍计数模型来应对未降雨的情况,但尝试后发现准确率极低,只有 10% 左右,整理数据后分析得到数据的趋势具有连续变量的特征,趋势上大致更加符合回归模型这一类.

随后扩大研究的范围,添加了云层覆盖,湿度,风速为新的自变量,由于此时线性回归在多个自变量的相互影响下适应性较差,转变模型,根据深度学习的基本原理和神经网络中防止过拟合的技术,运用了神经网络回归模型进行拟合,在仅使用 5 个神经层,共 200 多个神经元的情况下取得了绝对平均误差仅为 9mm 的优异结果. 同时,我们也展开了遗传算法训练网络的尝试,在有 4 层,分别有 5, 20, 20, 1 个神经元的网络中,发现每个个体网络无论输入,输出值均为 26mm 上下,同时,各个个体逐渐趋同演化,丧失了物种多样性,经过 100 代后,每个个体的适应度已相同,很难进化,最后无奈放弃此方法,猜测是遗传算法并不适合此种情况神经网络的训练,但依旧提供了一种解决问题的新思路.

在数据具有明显季节性质,且相邻的日期可能相互影响的情况下,我们也对时间序列模型进行了一定的尝试和研究,最后得到的效果也不如神经网络回归模型,原因可能是相互影响因素已经体现在了自变量中. 经过进一步的分析,尽管数据间并非存在明显的非线性关系,但是由于数据具有取值范围相对局限的特点,我们尝试了决策树模型,最终得到了满意的结果,绝对平均误差降到了 2.6mm,正确率飙升到 88% 左右. 尝试和分析过后,最后认为该决策树模型更加准确.

该模型具有预测率较高,和可以轻易扩展到新的自变量的优点. 生成该模型的方法可以轻易地被用于生成任何局部降雨模型,非常简单易行. 在考虑到气象的周期性而忽略人为影响的情况下,本模型具有较为准确的预测能力. 同时,实际上,降雨量的影响因素更加复杂,采用该模型时可以增加自变量的数量,进一步提高模型的准确度,具有一定的现实指导意义.

关键词: 神经网络回归模型 线性回归模型 遗传算法 决策树模型

¹以下简称为“气压”

2 问题重述

在过去的几十年中,随着各行各业的迅速发展,气象对于人类生产生活的影 响日益显著,对于气象预测的精确程度需要愈来愈高,降雨作为人们出行和农业生产的客观需要是非常关键的一部分.在当今时代,由于人类活动的影响日益增大和气象全球性变化的影响下,气象的预测越来越有必要与时俱进,因地制宜的适应局部气候的变化.

请以气温、降水、气压、云层覆盖、风速、湿度为自变量,预测气象的降雨量.

3 模型假设和符号说明

3.1 模型假设

该模型假设气象活动完全是周期性的,不考虑气候的逐渐变化,也不考虑人为因素的气候的干扰.

同时,假设在自然条件下,降水量只与气温、降水量、气压、云层覆盖、风速、湿度有关系.

3.2 符号说明

为了方便表述,我们将符号简写为 p, r, t, f, h, c , 同时,为了简化单位,我们规定了每个符号的默认单位.

以下若无特殊说明,则默认单位为下表所示的单位.

表 1: 符号说明

符号	说明	单位
r	降雨量	mm
t	温度	0.1°C
f	风速	0.1m s^{-1}
h	湿度	0.1%
c	云层覆盖	octas
p	气压	0.1hPa

4 模型建立与求解

4.1 获取数据

本文主要数据从欧洲气象网站² 关于 11 号 (STAI000011) 气象站 (一个奥地利气象站) 中获取,考虑到尽可能减少人类活动的影响,本文主要取用了该奥地利气象站附近地区从 1877 年到 2000 的 r, t, f, h, c, p , 为了简单起见,并没有选择辐射等可能影响因素相对较小的变量.

4.2 数据预处理

获取完数据之后,按照 30 天为一组进行平均³,整理完毕后,画出各个自变量的图像如下:

²见附录 ??.

³遗传算法中采用 1 个月为 1 组进行平均

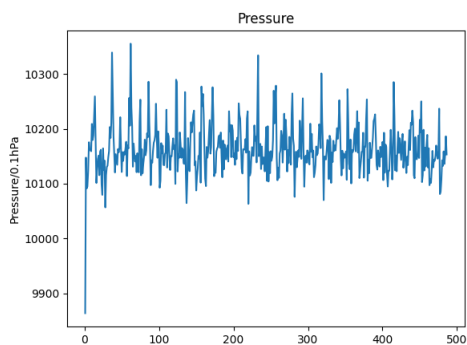


图 1: 气压覆盖关于时间的折线图

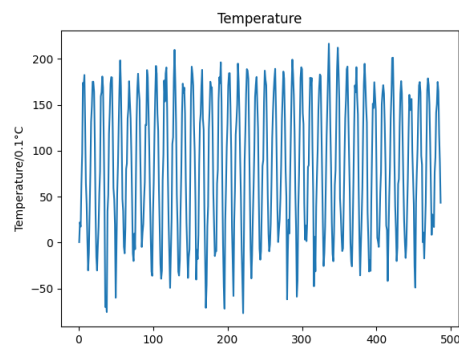


图 2: 气温关于时间的折线图

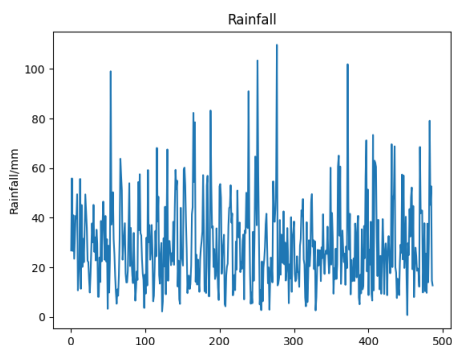


图 3: 降水量关于时间的折线图

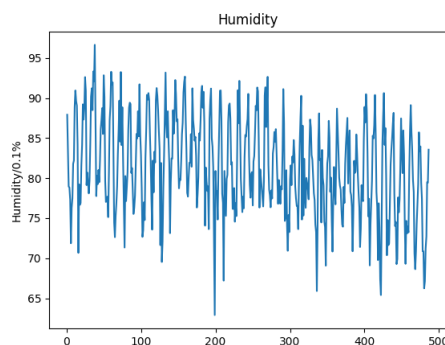


图 4: 湿度关于时间的折线图

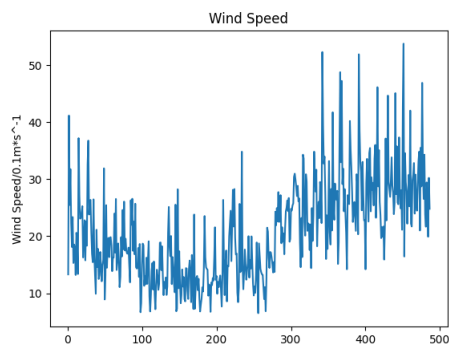


图 5: 风速关于时间的折线图

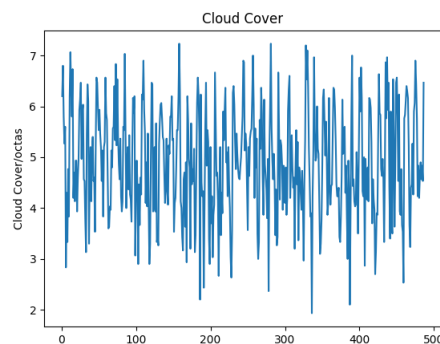


图 6: 云层覆盖关于时间的折线图

4.3 拟合模型

4.3.1 线性回归模型

简单研究, 发现 r, t, p 呈一定的周期关系, 且在周期内大致成趋势一致, 猜测之间很可能呈线性关系, 建立线性回归模型 $r = at + bp$, 使用最小二乘法让程序生成对应的模型⁴, 从而拟合最多的数据, 得到结果为

$$r = 0.1168736t + 0.00188574p$$

画出拟合图像与真实图像如图 ??.

可以看到数据趋势基本相符, 但是相比于实际数据没有那么极端, 此时正确率达到 52% 左

⁴该程序见附录 ??.

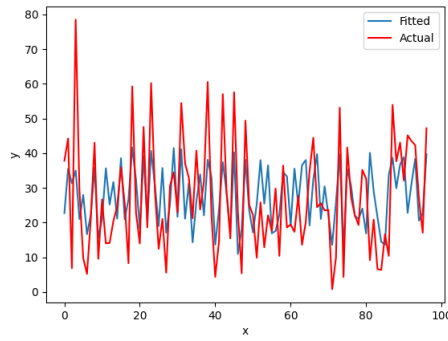


图 7: 拟合降水量和真实降水量的折线统计图

右, 绝对平均误差达到 10mm, 考虑到模型的吻合性其实不弱, 应当是气象影响因素比较复杂的缘故, 于是寻找新的自变量来考虑的更加全面. 经过一番对气象基本的研究, 我们添加了 f, c, h 这三个比较相关的变量.

根据对数据关系的分析, 发现此时已经有部分数据并不是时刻满足线性关系, 但总体上还比较吻合, 尝试继续用线性回归模型刻画, 使用最小二乘法, 得到模型为:

$$r = 6.264621458073717c + 0.23936949821555784h + 0.18358187666124123t \\ - 0.004092803931093655p + 0.23939598907957882f$$

该模型的描述能力确实有一定程度上的提高, 绝对平均误差保持在 10mm, 但是正确率达到了 64% 左右, 以下是 5 个自变量时的拟合图像:

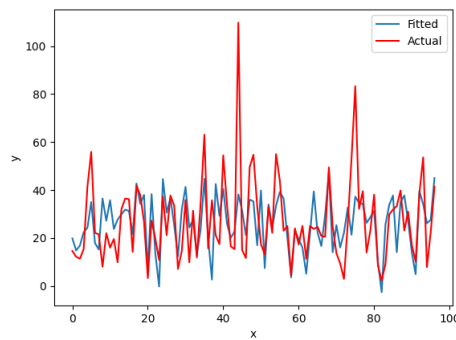


图 8: 拟合降水量和真实降水量的折线统计图

观察后可以发现该模型更加准确了, 但是离设想中还有一定的距离. 考虑到数据波动较大, 因此得出: 回归模型适合描述该问题的模型. 但线性回归模型局限性略大, 可能在变量之间的相互影响的作用下导致某些关系并非线性.

4.3.2 障碍计数模型

尝试障碍计数模型, 仅得到了 6% 的准确率, 因此障碍计数模型并不适合解决该问题.

4.3.3 神经网络

经过最初的尝试后, 首先确定了 5 个神经元层, 分别是 64,64,64,64,1, 若神经元数量再多过拟合的风险很大, 若过少也很容易造成对训练数据和测试数据正确率都降低的情况, 这样的数据

是比较合理的. 在对神经网络激活函数的选择中, 我们最初选择了`sigmoid`, 后来经过讨论, 认为该二分类激活函数不适合该背景, 改为了`leaky relu`函数, 大大提升了最初的训练速度.⁵

经过初步调整参数, 将训练次数定为比较合适的 300 次, 200 ~ 300 次均有不错的效果, 但此时过拟合已经十分严重, 在训练数据上达到了甚至达到了 100%, 然而在测试数据上只有糟糕的 60%, 经过一系列的尝试, 我们使用了L2正则化和`dropout` 的方式, 再加上早停的策略 (采用的是比较宽松的早停策略, 并且会记录下训练中的最佳权值, 训练一般进行 100 多次就会被早停), 成功的把训练数据和测试数据拉到一个水平线上, 由于该模型需要用几个 64×64 的矩阵描述, 过于占篇幅, 故在此处不给出.

此时, 模型已经达到了一个比较优秀的地步, 正确率达到了 74% 左右, 绝对平均误差降到了 8mm 左右, 绘制一下拟合图像:

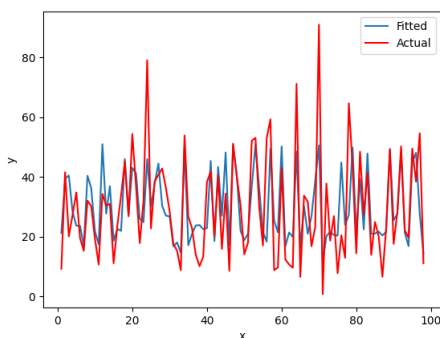


图 9: 拟合降水量和真实降水量的折线统计图

非常明显, 相比与之前的线性回归模型, 神经网络回归模型的拟合准确度明显大大增强, 但是对于某些特别极端的值似乎依旧有一些缺失, 不过总体拟合的还算准确.

猜测用遗传算法训练网络会有更好的结果, 我们也展开了尝试, 但经过切换各种激活函数和初始化方法, 发现网络得到的值方差较小, 有靠拢趋势, 在经过 1000 代后, 所有的个体的适应度都逐步收敛到一个值上, 同时所有个体的策略都是无论输入, 输出都是相同的值, 并不符合现实情况, 对现实指导的意义不大. 针对这种情况, 我们猜想是因为遗传算法对局部搜索能力不够强大, 导致了网络的学习效率低下. 同时我们采用的算法也不太适合用于神经网络模型训练, 我们只将其作为一种特殊的模型进行尝试, 提供新的思路.

4.3.4 决策树模型

下一步就是对模型类型进行进一步的尝试, 我们首先对与神经网络模型在处理关系不明显数据方面同样显著的决策树模型, 进行了检验, 从理论上来看, 数据间并非存在明显的非线性关系, 这样的问题应当不太适合使用决策树模型刻画, 但是对于降水量这样取值范围较为固定和小的特殊问题, 用决策树模型可能也会有特殊的好效果, 决策树模型的代码因为与线性前两个模型相似度很高, 在此处不贴出. 决策树模型的拟合图像如图 ??.

完美的拟合图像. 绝对平均误差从 8.9 骤降到 2.7, 而正确率飙升到 88%, 结果较好. 仔细分析, 可以得知也许是因为欧洲气象较为平均, 没那么极端, 决策树模型究其本质还是一种另类的分类讨论, 若是在气象多变的广东地区, 效果也许反而不如神经网络模型. 神经网络模型和线性回归模型从本质上探索他们之间的数量关系, 但决策树模型只是分类讨论罢了, 对于出现次数少

⁵该神经网络训练程序见附录 ??.

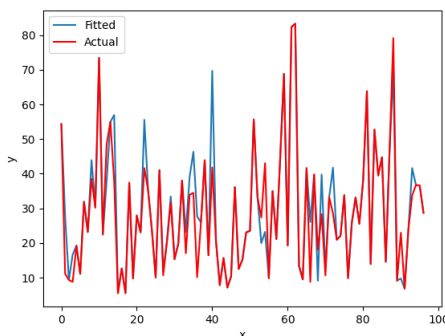


图 10: 拟合降水量和真实降水量的折线统计图

的预测能力不足,但是实践证明,也是一种非常切实可行的方案了.⁶

5 模型的优缺点与改进方法

5.1 线性回归模型

基于线性回归模型在仅研究几个变量的时候准确度和神经网络回归模型相差无几,具有操作十分简单,预测能力也不算很弱的特点,改进方法也许可以采用并非完全线性的回归模型,对于某些变量进行更加具体的分析,采用其他的基本函数进行描述效果可能会好一点

5.2 神经网络回归模型

基于神经网络的模型具有普适性更加强的特点,在降雨量受到如此多方面的影响的情况下对其中的几个关键变量进行研究能达到还不错的效果,实践起来也不算特别麻烦.

该方法可以利用神经网络强大的调整和适应能力用来研究各个地区的局部降雨量.不足之处在于,若是把时间范围拉得很大,会导致模型由于气候变化和人类活动导致不准确性增加,而且模型也并没有针对季节等进行描述.

可以从两方面考虑增强方法,一是结合一定的时间序列模型来解决气候变化和人类活动的问题,因为这两个因素也是按照一定的规律变化的,神经网络可以学习出其中的趋势;二是考虑季节性的特点,一定程度上可以使用有季节性的时间序列模型来缓解.但还有一种思路就是添加更多的自变量来间接反应季节的影响,因为气象总体是呈周期性的,例如可以添加风向作为自变量,来间接的反映季风,进而反映出季节的变化,感觉这种思路可能会好一点.

5.3 决策树模型

由于决策树模型具有可以随意调整叶子节点和非叶子节点的分界点,从而使数据尽可能的良好匹配所有的数据的特性和天气取值范围并不是很大的特点,使得它对于气候这种并非不具有显著函数关系的问题拟合程度非常高,几乎可以说是十分精确的预测了.

另一方面,对于极端情况的预测效果一定是很差的,因为数据被设计的尽量靠近正常的数据了,就会使得偏离不常出现的值.

⁶关于决策树模型的结果,由于仅能用大段文字分类描述,篇幅很大,故不放在附录中.

但是，若是我们把前两个模型综合起来看，使用线性回归模型和神经网络回归模型生成极端数据，再交由决策树模型拟合，就可以在一定程度上缓解这个问题，三个模型可以相辅相成，决策树模型也可以反过来帮助神经网络模型缓解过拟合的问题。

附录

.1 源代码

完整代码和数据

.2 数据来源

European Climate Assessment Dataset

.3 线性回归模型核心代码

```
1 data = pd.read_csv('merged_data_file.csv')
2 y_data = pd.read_csv("../tmprrr.csv")
3 y_data = y_data['RR'].values
4 data_numpy = data.to_numpy()
5 print(data_numpy)
6 data_without_first_column = data_numpy[:, 1:]
7 print(data_without_first_column)
8 # 创建线性回归模型
9 model = sm.OLS(y_data, data_without_first_column)
10 res = model.fit()
11
12 print("train rate:", get_rate1(res, data_without_first_column, y_data
13 ))
14 print("train rate2:", get_rate2(res, data_without_first_column,
15 y_data))
16 test_data_num = int(len(data_without_first_column) * 0.2)
17 test_idx = np.random.choice(len(data_without_first_column),
18 test_data_num, replace=False)
19 test_x = data_without_first_column[test_idx]
20 test_y = y_data[test_idx]
21 print("test rate:", get_rate1(res, test_x, test_y))
22 print("test rate2:", get_rate2(res, test_x, test_y))
23 print(res.params, )
24 draw(range(len(test_x)), test_y, res.predict(test_x))
```

.4 神经网络回归模型核心代码

```
1     scaler = StandardScaler()
2     X_scaled = scaler.fit_transform(data_without_first_column)
3     y_scaled = scaler.fit_transform(y_data.reshape(-1, 1))
4     X_train, X_test, y_train, y_test = train_test_split(X_scaled,
5                                                         y_scaled, test_size=0.2, random_state=42)
6     early_stopping = EarlyStopping(monitor='val_loss', patience=10,
7                                     restore_best_weights=True)
8     model = tensorflow.keras.models.Sequential()
9     model.add(Dense(64, activation='leaky_relu'))
10    model.add(Dense(64, activation='leaky_relu', kernel_regularizer=l2
11                    (0.01)))
12    model.add(Dropout(0.5))
13    model.add(Dense(64, activation='leaky_relu', kernel_regularizer=l2
14                    (0.02)))
15    model.add(Dropout(0.5))
16    model.add(Dense(64, activation='leaky_relu', kernel_regularizer=l2
17                    (0.02)))
18    model.add(Dropout(0.5))
19    # model.add(Dense(64, activation='relu'))
20    model.add(Dense(64, activation='leaky_relu'))
21    model.add(Dense(1))
22
23    model.compile(loss='mean_squared_error', optimizer='adam')
24    model.fit(X_train, y_train, epochs=300, batch_size=32,
25              validation_data=(X_test, y_test), callbacks=[early_stopping])
```
