|  |
| --- |
| **机器学习** |
| 下了多少雨？ |
| 主 研 人：赵翰宇  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2019/3/15 | A | 初稿 | 赵翰宇 |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 竞赛名 | 竞赛背景描述（50字以内） | 类型（分类/回归） |
| 2019/3/4 | How Much Did It Rain?（Kaggle） | 对于农业而言，了解特定油田的降雨量非常重要。极化雷达发射具有水平和垂直方向的无线电波脉冲，可以从两个方向的不同反射率推断雨滴的大小和水凝物的类型，生成每小时雨量计总量的概率分布 | 分类 |

目录

[1. 背景描述 4](#_Toc25575_WPSOffice_Level1)

[1.1 竞赛赛题描述 4](#_Toc5940_WPSOffice_Level2)

[1.2 评估指标描述 4](#_Toc12357_WPSOffice_Level2)

[2. 数据来源及描述性统计分析 5](#_Toc5940_WPSOffice_Level1)

[2.1 大赛数据来源 5](#_Toc22683_WPSOffice_Level2)

[2.2 数据的描述性统计 5](#_Toc16174_WPSOffice_Level2)

[2.2.1 数据基本情况描述： 5](#_Toc5940_WPSOffice_Level3)

[2.2.2 数据字段介绍： 5](#_Toc12357_WPSOffice_Level3)

[2.2.3 数据描述性统计 7](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 7](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 7](#_Toc32293_WPSOffice_Level2)

[3.1.1 方案一特征提取 7](#_Toc16174_WPSOffice_Level3)

[3.1.2 方案一模型设计与建立 8](#_Toc32293_WPSOffice_Level3)

[3.1.3 方案一结果、排名等 9](#_Toc29922_WPSOffice_Level3)

[3.1.4 方案一算法流程图 9](#_Toc4787_WPSOffice_Level3)

[4. 算法比较 9](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 9](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 10](#_Toc16174_WPSOffice_Level1)

[5.1 总结 10](#_Toc23198_WPSOffice_Level2)

[5.2 建模思路 10](#_Toc16619_WPSOffice_Level2)

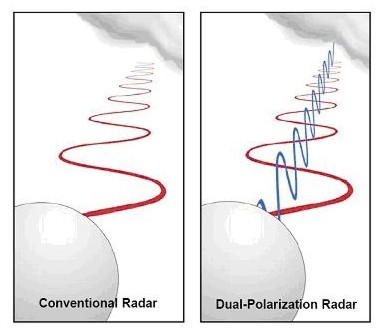
**1. 背景描述**

对于农业而言，了解特定油田的降雨量非常重要。然而，降雨在空间和时间上是变化的，并且不可能到处都有雨量计。因此，像雷达的遥感仪器能够提供广泛的空间覆盖。由于两个传感器的固有特性，从遥感观测中得出的降雨量估计值将永远不会与使用雨量计进行的测量完全匹配。目前，雷达观测通过使用附近的仪表来校正，并向需要知道下雨量的用户提供单一的降雨估算。本次比赛将探讨如何以概率的方式解决这个问题。

**1.1 竞赛赛题描述**

与传统的多普勒雷达不同，极化雷达（如图1-1）发射具有水平和垂直方向的无线电波脉冲。因为雨滴随着它们大小的增加而变得更平坦，并且因为冰晶倾向于垂直伸长，而液滴倾向于变平，所以可以从两者的不同反射率推断雨滴的大小和水凝物的类型以及取向。

在本次比赛中，选手通过在一小时内获得的极化雷达值和衍生数量，来生成每小时雨量计总量的概率分布。

图 1-1： 左图为多普勒雷达，右图为极化雷达

**1.2 评估指标描述**

获胜的标准是最小的持续排名概率分数CDF：

N 为数据集大小，z为实际记录的标准值（单位为mm）H（x）是Heavyside步长函数，即对于x≥0，H(x) = 1,其余则等于0。

如果有任意k满足结果为 ，则结果将被丢弃。

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

比赛的输入数值天气预报数据来自NOAA / ESRL全球集合预报系统（GEFS）重新预报版本第二版。培训数据包括2013年4月至11月的中西部玉米种植州收集的NEXRAD和MADIS数据。时间和位置信息已被删除，数据已经洗牌，因此不按时间或地点排序。测试数据包括同一月份但在2014年的相同雷达和仪表的数据。

这里是数据下载的超链接：https://www.kaggle.com/c/4195/download-all

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

train\_2013.csv  - 训练集，行数为ID，总共有113000个数据。列为18个特征。

test\_2014.csv  - 测试集

sampleSubmission.csv - 格式正确的示例提交文件

sample\_solution.py - 一个能够获取test\_2014.csv并生成sampleSubmission.csv的Python程序

**2.2.2 数据字段介绍：**

表2-1水文气象类型数据编码介绍

|  |  |
| --- | --- |
| **变量编号** | **变量含义** |
| **0** | 没有回声 |
| **1** | 中雨 |
| **2** | 中雨 |
| **3** | 大雨 |
| **4** | 雨、冰雹 |
| **5** | 大滴雨滴 |
| **6** | AP |
| **7** | 鸟群 |
| **8** | 未知 |
| **9** | 没有回声 |
| **10** | 干雪 |
| **11** | 湿雪 |
| **12** | 冰晶 |
| **13** | 软雹 |
| **14-** | 软雹 |

表2-2 train\_2013.csv数据表字段介绍

|  |  |
| --- | --- |
| **变量名** | **变量含义** |
| **TimeToEnd** | 雷达观测结束前的时间 |
| **DistanceToRadar** | 雷达和仪表之间的距离。该值经过缩放和舍入，以防止逆向工程测量仪的位置 |
| **Composite** | 垂直体积上的最大反射率 |
| **HybridScan** | 最接近地面的高程扫描反射率 |
| **HydrometeorType** | NSSL HCA中的九个类别之一。 |
| **Kdp** | 微分相 |
| **RR1** | 基于HCA算法的降雨率 |
| **RR2** | 基于Zdr算法的降雨率 |
| **RR3** | 基于Kdp算法的降雨率 |
| **RadarQualityIndex** | 从0（坏数据）到1​​（良好数据）的值 |
| **Reflectivity** | **反射率** |
| **ReflectivityQC** | 质量控制的反射率 |
| **RhoHV** | 相关系数 |
| **Velocity** | 多普勒速度 |
| **Zdr** | 以dB为单位的差分反射率 |
| **LogWaterVolume** | 有多少雷达像素充满了水滴 |
| **MassWeightedMean** | 平均墨滴大小，单位为mm |
| **MassWeightedSD** | 液滴尺寸的标准偏差 |
| **Expected** | 雨量计报告的实际降雨量 |

表2-3五种类型的“缺失数据”介绍

|  |  |
| --- | --- |
| **变量编号** | **变量含义** |
| **-99000** | 回波低于雷达的信噪比阈值。换句话说，真实值可能介于-14和-inf之间，但具体不知道。 |
| **-99901** | 范围折叠数据 |
| **-99903** | 由于光束阻塞或超出数据范围而未收集的数据 |
| **nan** | 导出的数量无法计算，因为某些输入是上述变量之一 |
| **999.0** | 无法计算RadarQualityIndex，因为像素位于回声的边缘 |

**2.2.3 数据描述性统计**

该竞赛数据数据量大，数据的缺失率高，大约85%的数据完全没有降雨，降雨量大于50毫米时，每类的样本数少于10个。因此数据筛选尤为重要。

3. 优秀算法思路

**3.1 方案一**

**3.1.1 方案一特征提取**

原始数据集中的每个变量都被扩展，以便每个扫描都有自己的列。计算每行的错误代码数（-99903，-99900等），以及非错误代码功能的总数。然后用np.nan替换错误代码，并计算每行的各种描述性统计数据（忽略nan值）。对于水文气象类型，每行的每个代码的数量都要计算在内。使用这些方法创建了191个特征，用于所有数据子集。

此外，对于具有8-17个雷达扫描的子集和大于18个雷达扫描的子集，对timetoend进行排序，并分别将小时划分为10个和20个等长段。在每个段中，计算平均值并将其用作特征。这些特性大大提高了数据集的分数，超过17次扫描，但对于扫描次数较少的子集几乎没有效果。

虽然功能选择可以删除大约50%的功能而不会降低评分，但团队没有删除很多功能。对于只有一个雷达扫描的子集，所有具有相同值的列都被删除。另外，由于距离太远，加上混合扫描的-99903和-99900错误码的计数，当包含70mm以上的标签时，分数急剧下降。去除大量雨水样本后，这些特征只略微降低评分，但仍被排除在外。

**3.1.2 方案一模型设计与建立**

方案一的整体建模方法是将预测视为从训练集估计的CDF的线性组合。根据分类算法给出的CDF相关标签的类概率对每个组件CDF进行加权。

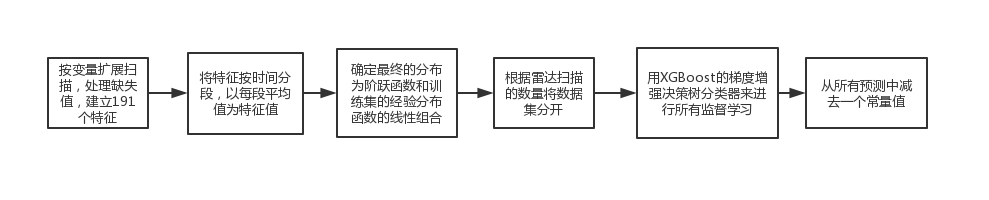
对一个实际降雨量为0毫米的样本，最好的预测是所有样本中的一行，而对于1毫米的实际降雨量，最好的预测是所有样本，除了0毫米位置的一个零点。利用这一点，如果分类器预测p（rain=0mm）=0.5和p（rain=1mm）=0.5，那么最好的CDF将是所有CDF的平均值，除了零（[0.5,1,1.5，…]）。

基于这种对数据的提前了解，方案一最终的分布为阶跃函数和训练集的经验分布函数的线性组合。数据集中存在相当数量的类不平衡。大约85%的数据完全没有降雨，降雨量大于50毫米时，每类的样本数少于10个。为了解决这个问题，方案一将高降雨量汇总为一个类标签。然后，方案一使用从培训数据中得出的降雨量分布来估算CDF。在做特征工程的时候，方案一决定根据雷达扫描的数量将数据集分开。对于一个20次雷达扫描的样本来说，不可能像一次只有1次雷达扫描的样本那样具有相同类型的特征。拆分数据集减少了培训时间，因为发现大多数数据的功能更少。此外，解决这个问题还缓解了数据集的一些类不平衡问题，这些数据集具有更多的扫描。在只有一次扫描的数据集中，95%的数据有0毫米的降雨量，很少有大降雨量的数据。这允许对只有3个标签的数据集进行建模。相比之下，17次以上扫描的子集，只有48%的数据有0毫米的降雨量。对于这一套，方案一可以使用12种不同的标签。

方案一用XGBoost的梯度增强决策树分类器来进行所有监督学习。参数是可能防止过度拟合的大致估计值。

最后，方案一添加了一些后期处理，使分数提高了0.00001。对于每个数据子集，方案一从所有预测中减去一个常量值。减去的数字与训练集中大于70的雷达扫描比例大致相同。团队不确定这是一个巧合还是因为大量降雨增加了分数，从每个预测中减去一点点，就减轻了他们的影响。

**3.1.3 方案一结果、排名等**

 此方案获得了第一名，公共排行榜上C为0.00752，私人排行榜上C为0.00748，分别排名第一位。

**3.1.4 方案一算法流程图**

**图3-1**

4. 算法比较

**表4-1 算法比较**

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **CDF** | **根据雷达扫描的数量将数据集分开** | **XGBoost的梯度增强决策树分类器** | **Python的Sklearn** |

个人认为算法1的特征处理非常合适。在数据缺失率很高，特征很多的情况下，算法1 没有像其他算法一样舍去大量的缺失数据，对于数据特征集中的问题，它将数据集分类散开，分段取每段平均作为特征，相当于用多个特征来描述一个变量的作用。同时，它经过测试，将一小部分影响预测的数据删除。

通过对数据的观察，算法1建立的分布模型为线性组合模型，拟合度较好。因此训练后得到的结果较好。

5. 总结与展望

**5.1 总结**

此次下雨分布的预测竞赛的关键在于分布模型的建立和高缺失率，且集中性很强的数据的处理，设计的算法以梯度增强决策树分类算法为最优。

**5.2 建模思路**

在特征提取上，我会将特征值划分出来，以每列特征的平均值进行填充，同时将每个雷达的预测值取平均作为特征值。在分布的模型上会选择经验分布模型。算法依然选择XGBoost决策树算法进行训练，最后再根据测试集进行数据调整。