|  |
| --- |
| **机器学习** |
| AMS 2013-2014 太阳能预测竞赛 |
| 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** |
| 2019/2/27 | 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） |
| 2019/2/14 | AMS 2013-2014 Solar Energy Prediction Contest（Kaggle） | 电力公司需要对太阳能源生产进行准确的预测，以便在可再生燃料和化石燃料之间取得适当的平衡。利用统计和机器学习技术与数值模型结合使用，以产生更精确的预报。 | 回归 |

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 5](#_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： 6](#_Toc12357_WPSOffice_Level3)

[2.2.3 数据描述性统计 8](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 8](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 8](#_Toc32293_WPSOffice_Level2)

[3.1.1 方案一特征提取 8](#_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)

[3.2 方案二 9](#_Toc29922_WPSOffice_Level2)

[3.2.1 方案二特征提取 9](#_Toc16174_WPSOffice_Level3)

[3.2.2 方案二模型设计与建立 10](#_Toc32293_WPSOffice_Level3)

[3.2.3 方案二结果、排名等 10](#_Toc29922_WPSOffice_Level3)

[3.2.4 方案二算法流程图 10](#_Toc4787_WPSOffice_Level3)

[3.2 方案三 10](#_Toc29922_WPSOffice_Level2)

[3.2.1 方案三特征提取 10](#_Toc16174_WPSOffice_Level3)

[3.2.2 方案三模型设计与建立 10](#_Toc32293_WPSOffice_Level3)

[3.2.3 方案三结果、排名等 10](#_Toc29922_WPSOffice_Level3)

[4. 算法比较 10](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 11](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 12](#_Toc16174_WPSOffice_Level1)

[5.1 总结 12](#_Toc23198_WPSOffice_Level2)

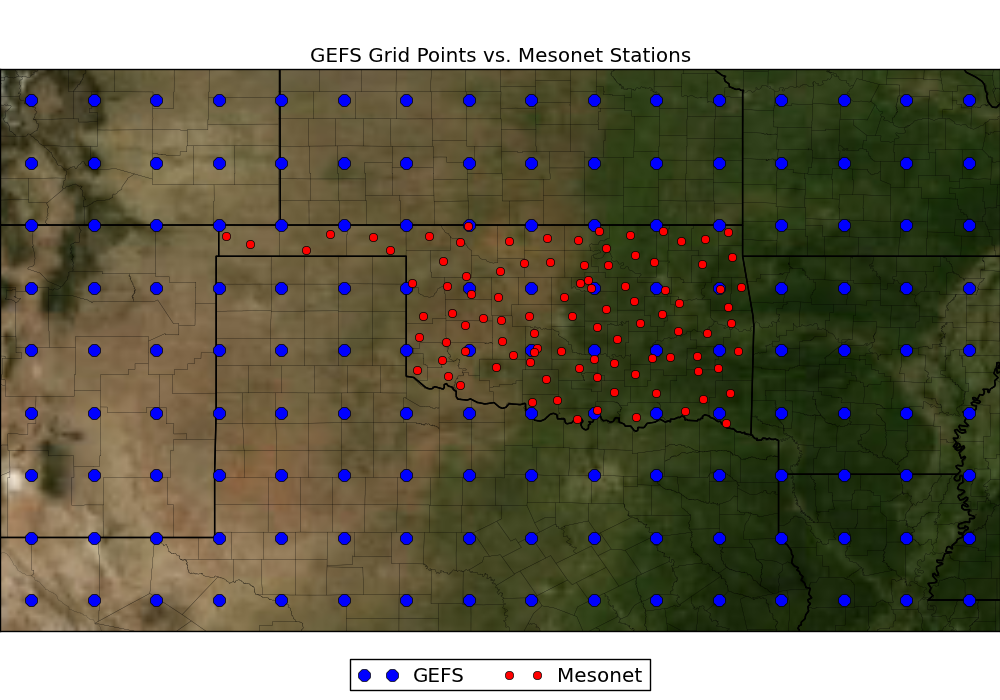
[5.2 建模思路 12](#_Toc16619_WPSOffice_Level2)

1. 背景描述

太阳能和风能等可再生能源比化石燃料发电提供了许多环境优势，但它们产生的能量随着天气条件的变化而波动。电力公司需要准确预测能源生产，以便在可再生能源和化石燃料之间取得适当的平衡。预测中的错误可能导致公用事业公司因过度燃料消耗或从邻近公用事业公司紧急购买电力而产生大量费用。功率预测通常源自数值天气预报模型，但统计和机器学习技术越来越多地与数值模型结合使用以产生更准确的预测。

**1.1 竞赛赛题描述**

本次比赛是发现哪种统计学和机器学习技术能够为太阳能生产提供最佳的短期预测。参赛者将预测98个俄克拉荷马州Mesonet站点的每日太阳能总量，这将作为比赛的“太阳能发电场”。Mesonet站点相对于GEFS数据的位置如图1-1所示。训练数据将来自1994 - 2007年。公共测试数据将来自2008 - 2009年。最近一段时间的私人测试数据将用于最终评估。

图 1-1： Mesonet站点相对于GEFS数据的位置

**1.2 评估指标描述**

参赛者将在test.csv文件中指定的每一天提交98个俄克拉荷马州Mesonet站点的每日太阳辐射总量的预测。平均绝对误差（MAE）是用于本次比赛的指标。它常用于回归问题和可再生能源行业比较预测性能。公式由下式给出：

与Root Mean Squared Error不同，MAE不会过度惩罚极端预测。

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。数据包括所有11个集合成员和预测时长12,15,18,21和24小时。训练数据将来自1994 - 2007年。公共测试数据将来自2008 - 2009年。最近一段时间的私人测试数据将用于最终评估。

这里是数据下载的超链接：https://www.kaggle.com/c/3354/download-all

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

sampleSubmission.csv文件是输出样例。

**train.csv**包含自1994年1月1日到2017年12月31日连续运行的98个俄克拉荷马州Mesonet站点（J m-2）的每日总入射太阳能。太阳能能量直接通过每5分钟Mesonet站点的日射强度计测量并根据所列日期的日出时间到23:55UTC来累计。

**station\_info.csv为99×4的列表，4列为站点名称，**纬度，经度和海拔（米），行是98个站点名称。

**gefs\_train.tar.gz**和**gefs\_train.zip**包含所有GEFS训练数据，每个压缩包有15个特征表格，每个表格有多维数组。第一个维度是模型运行的日期，它直接与train.csv文件中的每一行对应。第二个维度是要预测的11个GEFS成员。第三个维度是预测时长，以3小时为增量，从时长12小时一直增加到时长24小时。所有模型运​​行时间都从00 UTC（世界标准时间，以伦敦时间为标准）开始，因此尽管每年的当地太阳时间会有所不同，在此题中的时间将始终对应于相同的通用时间。第四维和第五维是纬度和经度统一的空间网格。文件中的经度在本初子午线的正度数上，因此从它们中减去360会将它们转换为与station\_info.csv类似的值范围。如图1所示。

**gefs\_elevations.nc**是一个netCDF4文件，其中包含GEFS网格点的模型海拔。由于模拟地形与现实世界相比是平滑的，因此特定纬度点的真实海拔可能与模型中的海拔不匹配。该文件包含两个海拔变量。所述elevation\_control变量包含该GEFS控制运行，这是第一个GEFS成员的海拔。所述elevation\_perturbation变量包含该GEFS扰动，是其他GEFS成员的海拔。

**2.2.2 数据字段介绍：**

表2-1 sampleSubmission.csv数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **Date** | 测试集日期 | 日期 | 0% |
| **其余变量** | 98个站点的总入射太阳能能量 | 离散 | 0% |

表2-2 **train.csv**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **Date** | 训练集日期 | 日期 | 0% |
| **其余变量** | 98个站点的总入射太阳能能量 | 离散 | 0% |

表2-3 **gefs\_elevations.nc**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **elevation\_control** | 第一个GEFS成员的海拔 | 离散 | 0% |
| **elevation\_perturbation** | 其他GEFS成员的海拔 | 离散 | 0% |

表2-4 **15个特征**字段介绍

|  |  |  |
| --- | --- | --- |
| **变量名** | **变量含义** | **单位** |
| apcp\_sfc | 地表3小时累积降水量 | kg m-2 |
| dlwrf\_sfc | 在地表的向下的长波辐射通量的平均 | W m-2 |
| dswrf\_sfc | 在地表的向下的短波辐射通量的平均 | W m-2 |
| pres\_msl | 平均海平面的气压 | Pa |
| pwat\_eatm | 可沉淀的水在整个大气层深处 | kg m-2 |
| spfh\_2m | 地面以下2米的特定湿度 | kg kg-1 |
| tcdc\_eatm | 整个大气深度的总云量覆盖 | % |
| tcolc\_eatm | 整个大气中的总柱浓缩冷凝液。 | kg m-2 |
| tmax\_2m | 过去3小时的高于地面2米的最高温度。 | K |
| tmin\_2m | 过去3小时的高于地面2米的最低温度。 | K |
| tmp\_2m | 高于地面2米的当前温度 | K |
| tmp\_sfc | 地表温度 | K |
| ulwrf\_sfc | 在地表向上的长波辐射 | W m-2 |
| ulwrf\_tatm | 在大气顶部向上的长波辐射 | W m-2 |
| uswrf\_sfc | 在地表向上的短波辐射 | W m-2 |

表2-5 **gefs\_train.zip中的**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **time** | 对应train.csv日期的入射太阳能能量 | 离散 | 0% |
| **lat** | 纬度值 | 离散 | 0% |
| **log** | 经度值 | 离散 | 0% |
| **ens** | 集合成员 | 离散 | 0% |
| **fhour** | 持续时间 | 离散 | 0% |

**2.2.3 数据描述性统计**

该竞赛数据维度较高，除去时间维度，特征包含15个天气特征×5个时间段特征，同时要建立对应11个GEFS成员的模型。但是数据的缺失率低，可以得到较好的训练模型。对于数据的集中性和离散性分析，该竞赛包含不同经纬度的地区的关联度不高的特征的数据，因此分析数据集中性和离散性意义不大。

3. 优秀算法思路

**3.1 方案一**

**3.1.1 方案一特征提取**

此方案的模型使用预测文件的所有功能而不应用任何预处理，因此我们将所有的75个预测作为功能

对于每个站点，方案选取了4个最近的站点的预测值，这4个站在中心站周围形成类似正方形的区域。对于这些站点中的每一个，都取了5个时间段值。所以此方案一共取有4个站点，每个Mesonet站点每天有5个时间段特征值（4x5=20个值）。因为它有15个特征（15个netcdf文件），特性的数量乘以15，所以在每个模型中，每个Mesonet站点每天有300个特征。

除了这些特性之外，我们还使用了特征：月份、站点ID、每个站点经纬度差异以及中心站点到周围4个站点中每个的距离。

**3.1.2 方案一模型设计与建立**

首先，方案团队先进行了一些基准测试来了解数据集的表现。

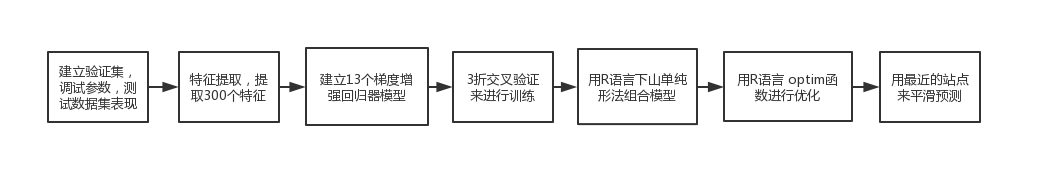
方案一开始使用1000个估算器（或树）。在python中，设最大深度为5或7，之后进行网格搜索以找到学习率（以及python中的最大特征）。其余保留默认值。在此之后使用3000棵树，并重新调整学习率。它接近线性，估算数量增加三倍，学到的数量减少到三分之一。但由于调参太过耗时，团队决定用组合的方法去找最优解。最终发现不需要进行任何功能工程，并且使用python的梯度增强回归器（gbr）方法得到了最好的结果。

## 训练数据集时，方案使用了组合的方法，使用连续周期的3折交叉验证（3折分别是1994-1998,1999-2003,2004-2007年）对13个模型进行了训练。方案没有花太多时间构建复杂的特征，而是花时间在数据集上进行转换。最后一个模型是13个不同模型的组合，其中11个模型用于11个GEFS成员的预测，一个模型用于预测的中位数，另一个模型用于每天预测变量的总和。之后使用R语言中的Nelder-Mead下山单纯形法来进行模型组合。这样做是为了使训练数据集类似于测试数据集，并避免过度拟合。用于生成模型的工具有：R和python。在模型训练之后，我们使用R语言中的 optim函数进行了10倍的优化。最后做了另一个优化，使用最近的站点来平滑预测。

## 通过检验，组合算法似乎缩小了预测。因此最后将预测值乘以1.015，得到公共和私人排行榜的最佳测试结果。

**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MAE为1878228.94，私人排行榜上MAE为2107588.17，分别排名第一位。可能由于数据数值较大，误差值数值也较大。

**3.1.4 方案一算法流程图**

**图3-1**

**3.2方案二**

**3.2.1 方案二特征提取**

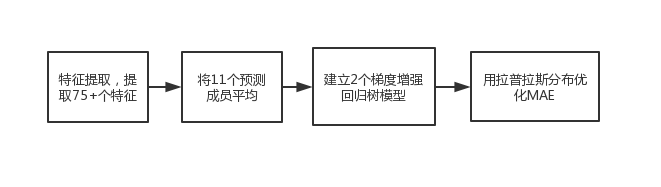
除了dswrf特征在5个时间点的线性组合外，此方案也没有任何的特征工程。方案直接提取15×5=75个特征，再加上一年中的每日和经纬度特征。

**3.2.2 方案二模型设计与建立**

方案先将11个GEFS预测成员平均化，之后用R语言的GBM包，建立了两个梯度提升回归树模型，模型中的分布参数选用拉普拉斯分布来优化MAE值。两个模型中，一个是基于近似内核的75个特征，另一个是基于最接近的4个特征的75个特征的加权平均值。

**3.2.3 方案二结果、排名等**

此方案获得了第三名，公共排行榜上MAE为1926384.23，排名第五位，私人排行榜上MAE为2158525.70，排名第三位。

**3.2.4 方案二算法流程图**

**图3-2**

**3.3方案三**

**3.3.1 方案三特征提取**

此方案在方案二的基础上，添加了dswrf和pwat这两个重要特征的派生特征。添加了月份，海拔，经纬度的相关性。最重要的是，对四个最近的GEFS站点进行了线性插值（按距离加权）

**3.3.2 方案三模型设计与建立**

此方案在模型设计上和方案二一致，算法用C#编写，优化方法不明。

**3.3.3 方案三结果、排名等**

此方案获得了第二名，公共排行榜上MAE为1905920.10，排名第三位，私人排行榜上MAE为2128115.99，排名第二位。

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** | **MAE** | **无** | **梯度提升回归器（gbr）** | **Python的Sklearn**  **R的optim** |
| **算法2** | **MAE** | **无** | **梯度提升决策树（gbm）** | **R的gbm** |
| **算法3** | **MAE** | **无** | **梯度提升决策树（gbm）** | **C#的gbm** |

个人认为算法1，2和3都运用了相同的算法，但在总体思路上不一致。在对11个预测成员的方法上，算法一用组合模型的大思路，对每个成员分别建立模型，训练后再取平均进行组合，这样的做法缺点是复杂度高，速度慢，而且最终结果缩小了预测，需要手动调参进行放大，问题可能出自于组合方法，应当找到使11个预测成员关联度更高的组合方法。优点是针对11个对象的预测较准，回归拟合度高。而算法二和三则是先对11个预测成员取均值，再进行回归训练，因此对预测成员的拟合度较低，且更缺乏针对性的意义。

除大体思路外，算法二在特征提取上的优点是将重要特征dswrf根据时间点进行了线性组合，算法三比算法二做的更好，增填了派生特征，并且增添了其余特征的相关性，对四个站点进行了按距离加权。而算法一没有太多关注特征本身的特性。但算法一同算法三一样，利用平均周围的特征来描述中心站点特征，增强了每个经纬度点的关联性，同时增添了大量的特征。因此特征丰富，利于模型学习。

5. 总结与展望

**5.1 总结**

此次太阳能预测竞赛的关键在于处理多成员多维数据的相关性，设计的算法以梯度提升回归算法为最优。对于多成员的处理，可以采取组合模型或者取成员均值等办法，但需要提高成员之间的相关性。最终结果的MAE值仍然较大，说明模型仍存在可提升的地方。

**5.2 建模思路**

个人认为，在特征提取上，提取15×5=75个特征值，同时进行75个特征值与经纬度特征与时间特征的相关性分析。在算法上，改进GBM算法，利用XGBoost算法进行训练，预测效果会更佳，GBM在优化时只用到一阶导数信息，xgboost则对代价函数进行了二阶泰勒展开，同时用到了一阶和二阶导数。优化效果更加。且xgboost在代价函数里加入了正则项，能够控制模型的复杂度。因此能减少训练时间。