节

摘要

关键字： 贪心算法　变步长搜索　 坐标变换

一、 问题重述

1.1 问题 1

1.2 问题 2

1.3 问题 3

二、 问题假设

- 1：假设天体 S 离 fast 射电望远镜足够远，射电望远镜口径与此距离相比可以忽略不计，此时来自天体 S 的电磁波可视为平行射入。

三、 符号假设

符号	意义
h	抛物面顶点到基准球面间的距离

四、 模型建立

为了预测用户是否遭到的电信诈骗，该问题本质上是一个逻辑回归问题，常用的模型有 logistical 回归，线性回归 (Linear Regression)，支持向量机 (SVM) 等。由于单个模型的泛华能力较差，为了获得更高的召回率，准确率，我们这里采用基于模型融合 (Stacking) 的方法来求解该逻辑回归的问题。

所谓模型融合，在训练数据上训练并使用多个模型进行预测，得到多组预测结果，也就是超特征，使用一个新的模型，对这些超特征再进行训练，训练一个从超特征到真实值的模型, 再将测试数据的超特征输入这些模型，得到最后的结果。(**??**) 这里我们采用多层模型融合方法, 如 (**??**) 所示的模型融合架构, 逐层叠加由当个模型加权组合而成的模型来作为最后的模型。

对于第 N 层，组合模型，假设对于其中每个模型的输入为:$\vec{Input}_N$, 对应的权重为 $W_{Input,N}$, 输出预测结果为 $\vec{P_{i,N}}$, 其中 i 为该层的所有模型的编号 $\vec{W_{i,N}}$, 其对应的权重为。则第 N+1 层的输入为: $\sum_{\forall i} \vec{P_{i,N}} \bullet \vec{W_{i,N}} + \vec{Input}_N W_{Input,N}$, 并且我们将数据集分为 k 块，对不同的模型进行 k 折交叉验证。通过该操作可以减少方差和降低过拟合。

这里我们使用的多层模型融合具有 3 层，对于每一层融合模型，其输入为下一层的融合模型输出和单个型 (这里包括 RandomForestGini, WeightedEnsemble_L2(来自于第二层的融合模型的输出), RandomForestEntr, ExtraTreesGini, ExtraTreesEntr, XGBoost, CatBoost, NeuralNetTorch, LightGBMLarge, LightGBM, NeuralNetFastAI, LightGBMXT, KNeighborsDist, KNeighborsUnif)。

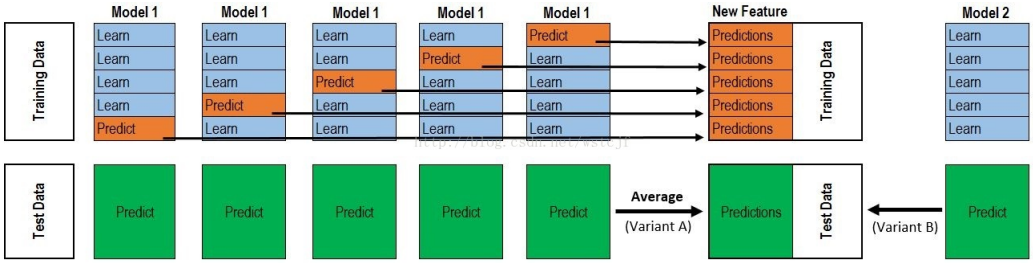

图 **1**　模型融合示意图

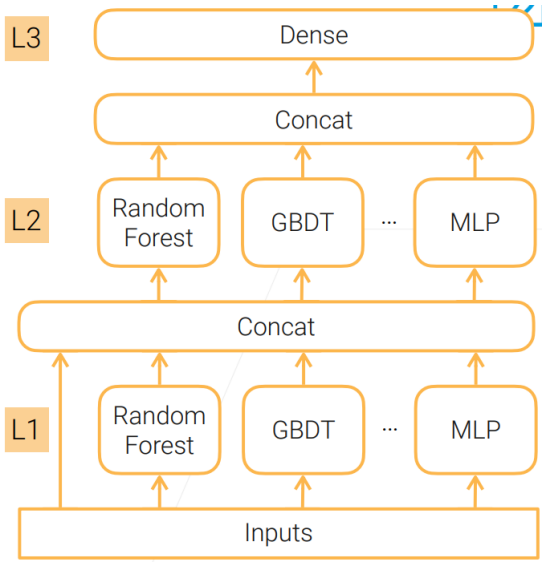

图 **2**　多层模型融合架构示意图

五、 模型求解以及结果分析

对于模型的求解，我们使用 python 进行求解。我们随机选取 70%(70 万) 的数据用作训练数据，将 20%(20 万) 数据作为测试集, 以及将剩下的 10%(10 万数据) 作为测试集。在进行 k 折交差差验证的训练后，我们的多层融合模型在测试集上取得了如下的效果:

- 准确率 (accuracy): 0.9999288888888889
- 平衡准确率 (balanced_accuracy): 0.9996162978031218
- 马尔萨相关系数 (mcc): 0.9995542858142105
- F1 分数: 0.9995931703472036
- 精度 (precision): 0.9999491281842577
- 召回率 (recall): 0.9992374658448243

我们分析了该多层融合模型的不同的单个模型在测试集/验证集的得分 (**??**), 从中可与看出，多数模型在验证集以及测试集上的得分 (准确率) 都是 99% 以上，只有模型 KNeighborsDist 和 KNeighborsUnif 的得分较低 (只有 93%) 左右，这可能是因为这两个模型都是使用 KNN 模型进行训练的，是基于每个样本的附近样本来预测该样本，从上文的分析中我们知道，由于样本十分不均衡，所以导致了 KNN 模型得到的结果较差，而在多层融合模型中，我们可以通过学习系数权重从而降低 KNN 模型的影响，所以并不会对结果产生较大影响。同时，我们也可以看出，相比于神经网络，随机森林 (RandomForestEntr)，XGBoost 等传统模型也可以取得不同的结果，这一点与论文 [] 的观点一致。

同时我们也分析了单个模型的在测试集，验证集上的预测耗时 (**??**)，以及训练耗时。

从中我们可以看到，对于所有模型，在验证集以及测试集上的预测耗时极低，基本可以忽略不计，这表明我们选取的模型在训练好进行预测时速度较快。但是从训练时间这一项可以看出：对于神经网络类的模型 (NeuralNetTorch, NeuralNetFastAI), 在进行时耗时较长，这也是符合一般经验的，训练神经网络需要在每个周期内都进行，梯度下降调参以及损失反向传播，而这两个操作都是比较耗时的。而对于传统的机器学习 (如随机森林，XGBoost)，其训练时间是可以忽略不计的，而且从上文中，我们知道，这种类型的模型也具有较高的精度。但是需要注意的是，在参数文件中，神经网络类的模型由于网络参数较多，其参数文件也比较大，大概是其他传统机器学习模模型的几个数量级倍。

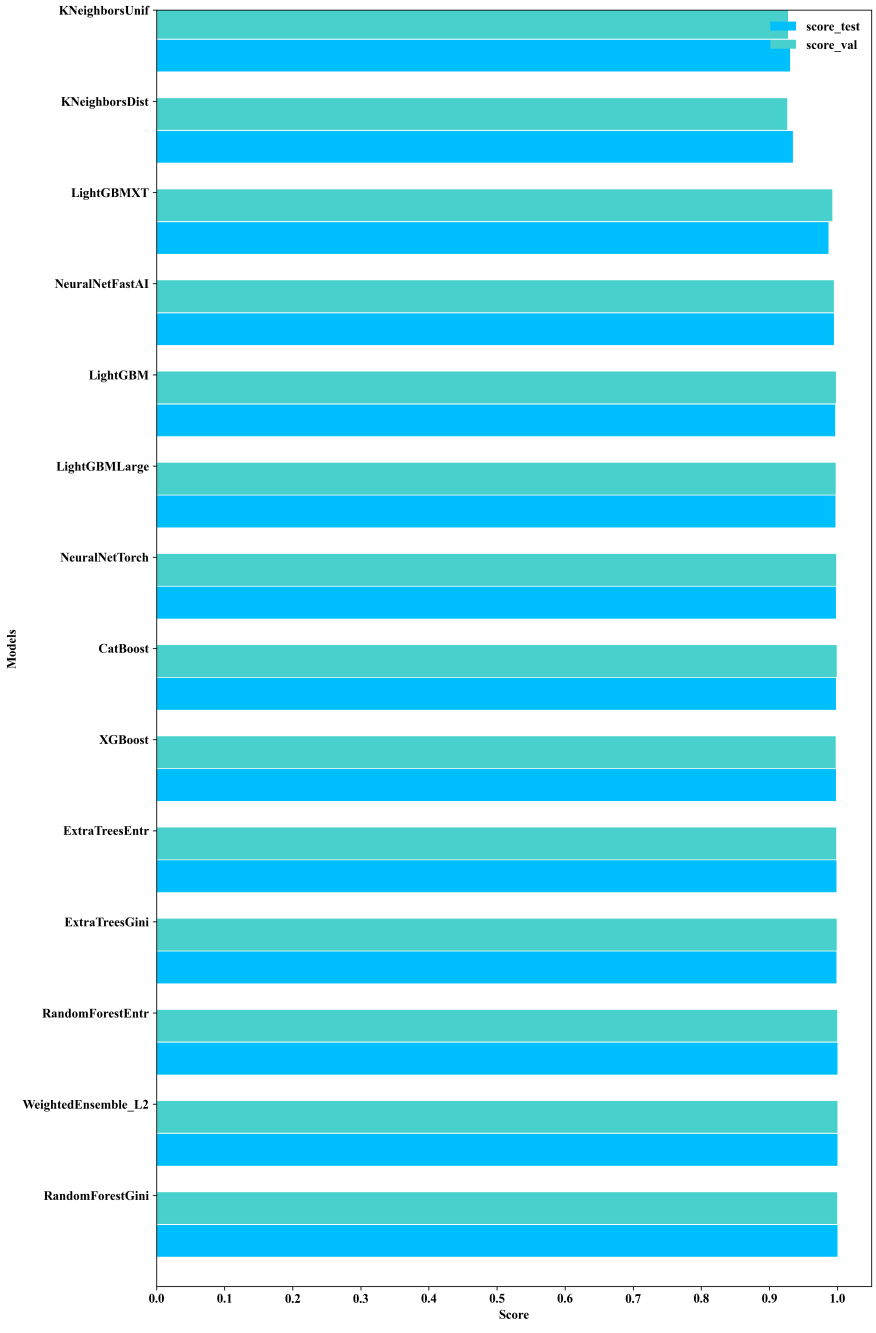

图 **3**　单个模型的验证/测试得分

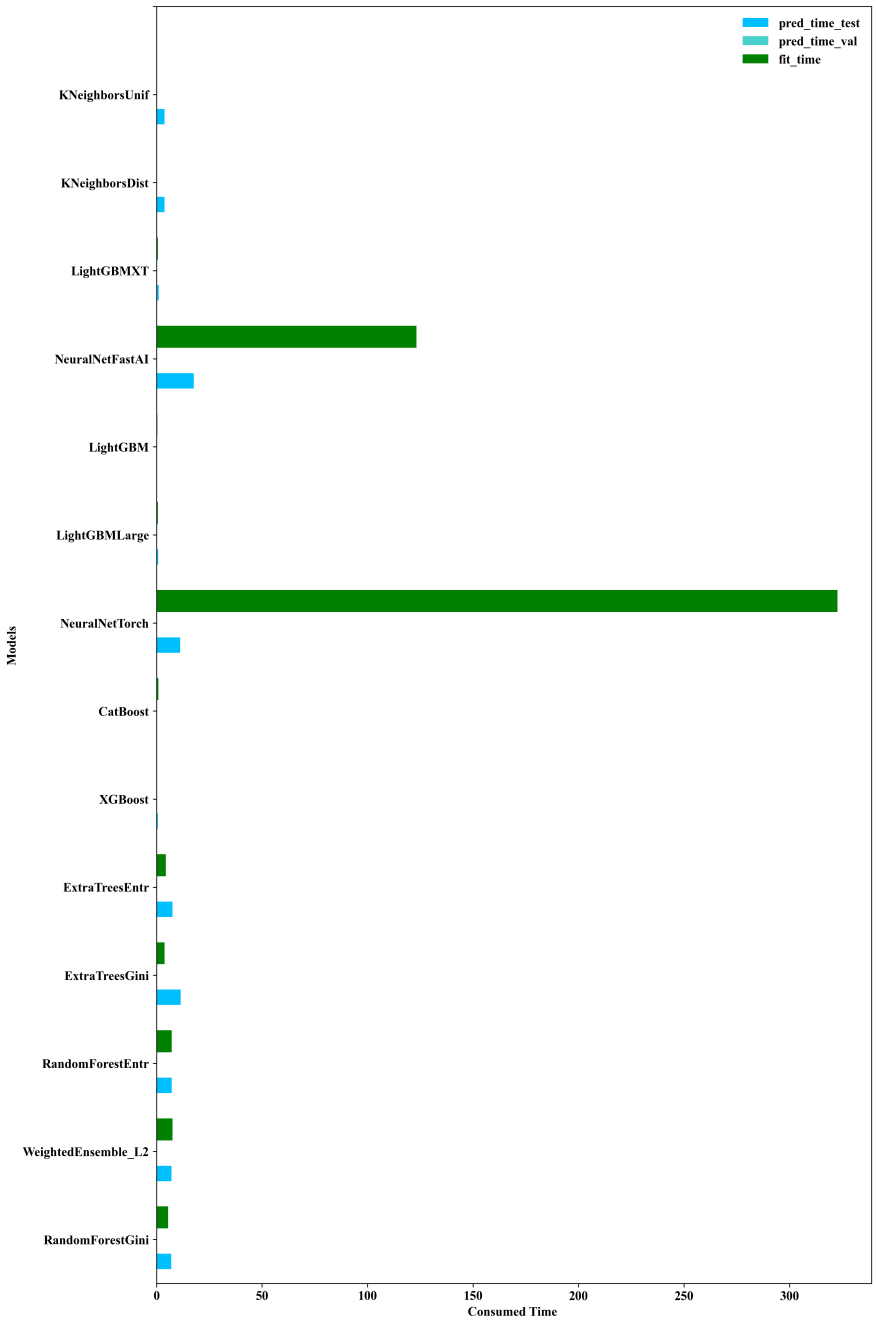

图 **4**　单个模型在训练/测试集的预测耗时以及训练耗时

模型的优化

六、 模型的评价与推广

6.1 模型的优点

1. 模型参考了”FAST” 有关的可靠文献，提出了较合理的衡量反射面板与理想抛物面误差的方案。

2. 模型将连续问题离散化，进行变步长搜索，先粗调再细调，搜索速度快，巧妙地解决了优化问题。

3. 模型将三维问题转化到二维平面进行分析，进行了相当的简化。

6.2 模型的缺点

1. 模型近似过程较多，对于光线光路可能存在一定影响。

2. 模型所得到的结果难以验证。

3. 忽略了馈源舱不同部分接受到的光强差异所产生的影响。

6.3 模型的改进与推广

- 可对于每个反射面板以球面反射进行更细致的分析，进一步提高接受比。
- 可采用遗传算法等搜索出贪心算法中开始调节的起点。
- 模型本身在”FAST” 球面射电望远镜已经大获成功的前提下具有一定的可行性, 在其他球面结构的望远镜中可同样适用，具有普遍意义。

参考文献

[1] 朱丽春. 500 米口径球面射电望远镜 (FAST) 主动反射面整网变形控制 [J]. 科研信息化技术与应用,2012,3(04):71.

[2] 李明辉, 朱丽春. FAST 瞬时抛物面变形策略优化分析 [J]. 贵州大学学报 (自然科学版),2012,29(06):25

[3] 钱宏亮. FAST 主动反射面支承结构理论与试验研究 [D]. 哈尔滨工业大学,2007：2(27).

附录 A　问题 1 的代码