中国移动 China Mobile

5G++创新实训基地

技术实操的练兵场 · 能力认证的人才站 · 5G应用的孵化器

版权声明

中国移动 China Mobile | 5G+创新实训基地

技术实操的练兵场 · 能力认证的人才站 · 5G应用的孵化器

基于 KPI 的数据挖掘建模

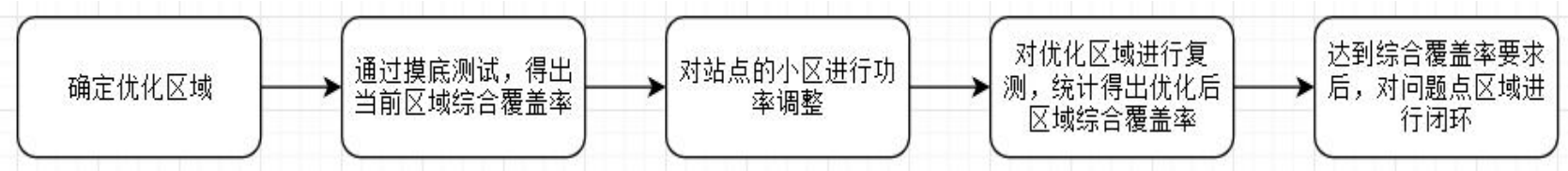
2020年11月

诺基亚-廖文哲

Contents

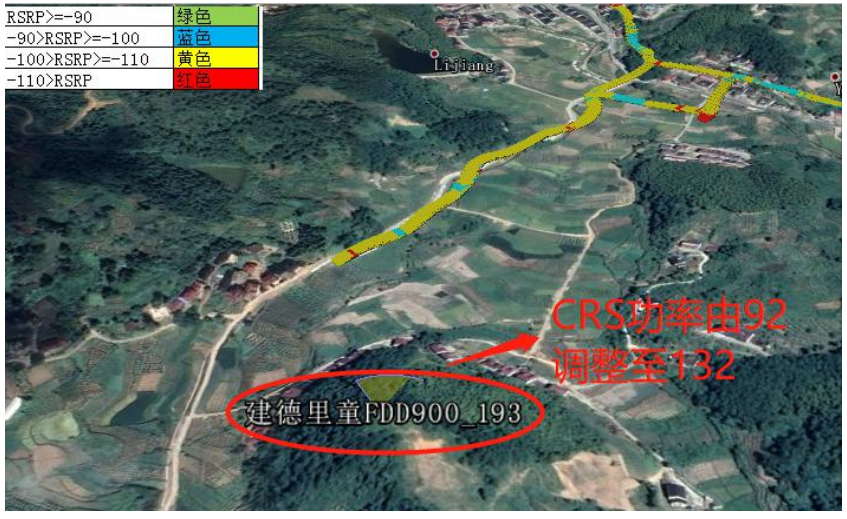
一.课题背景

由于4G终端的不断普及和业务量的持续增加，对日常网络覆盖提出了更高的要求，为不断提升不同场景的深度覆盖程度，主要通过提升功率、调整天线方位角、新建站点等方式进行解决。在完成相关提升深度覆盖的调整后，目前主要通过传统路测去现场获得深度覆盖数据，再由人工导出、统计数据进行效果评估。



实际上每进行一次大规模深度覆盖优化，涉及区域可能就会达到上千个栅格或簇等区域，且很多区域均处于偏远山村，这将耗费巨大的财力及人力评估优化效果，所以提供一种极简+智慧的新技术方案用以功率优化区域综合覆盖率评估迫在眉睫。

优化前



优化后

二、特征工程-数据获取

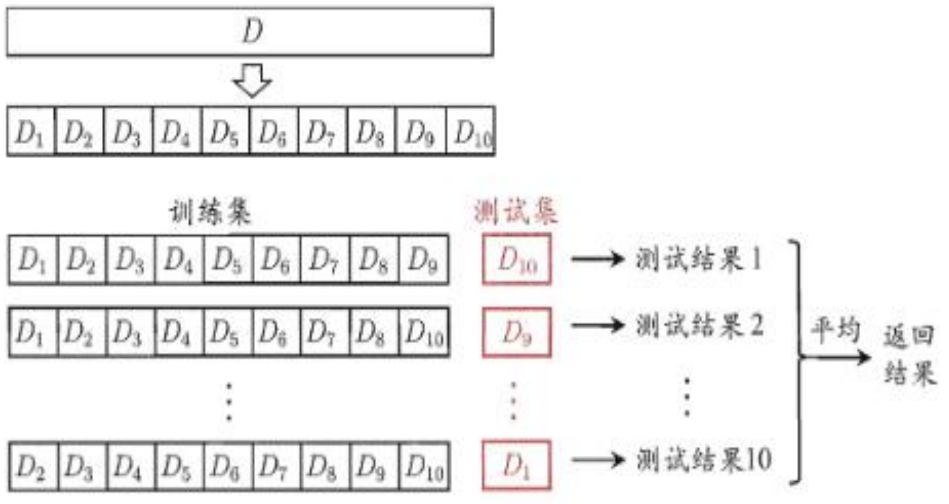
特征工程，是指用一系列工程化的方式从原始数据中筛选出更好的数据特征，以提升模型的训练效果。业内有一句广为流传的话是：数据和特征决定了机器学习的上限，而模型和算法是在逼近这个上限而已。由此可见，好的数据和特征是模型和算法发挥更大的作用的前提。特征工程通常包括数据预处理、特征选择、降维等环节。在数据获取阶段，首先根据业务专家经验获取本项目需要预测的标签值和特征值：

标签值	备注
覆盖提升率	覆盖率提升了多少

特征值	备注
摸底测试覆盖率(%)	综合覆盖率指标
栅格属性	用以区分区域的属性（道路或自然村）
（RSRP>=-110dBm&SINR>=-3dB）--覆盖率分子数	用以计算综合覆盖率
总采样点数--覆盖率分母数	
RSRP>=-110采样点数	
SINR>=-3采样点数	
最近站点距离	影响区域内综合覆盖率提升程度
站点位于栅格中心点的方位角	
1Km内升功率小区数量	
1Km内升功率幅值(求和)	
1Km内升功率幅值(均值)	
2Km内升功率小区数量	
2Km内升功率幅值(求和)	
2Km内升功率幅值(均值)	
平均RSRP	其他关联指标
平均SINR	
RSRP>=-105采样点数	
RSRP>=-100采样点数	
RSRP>=-90采样点数	
SINR>=0采样点数	

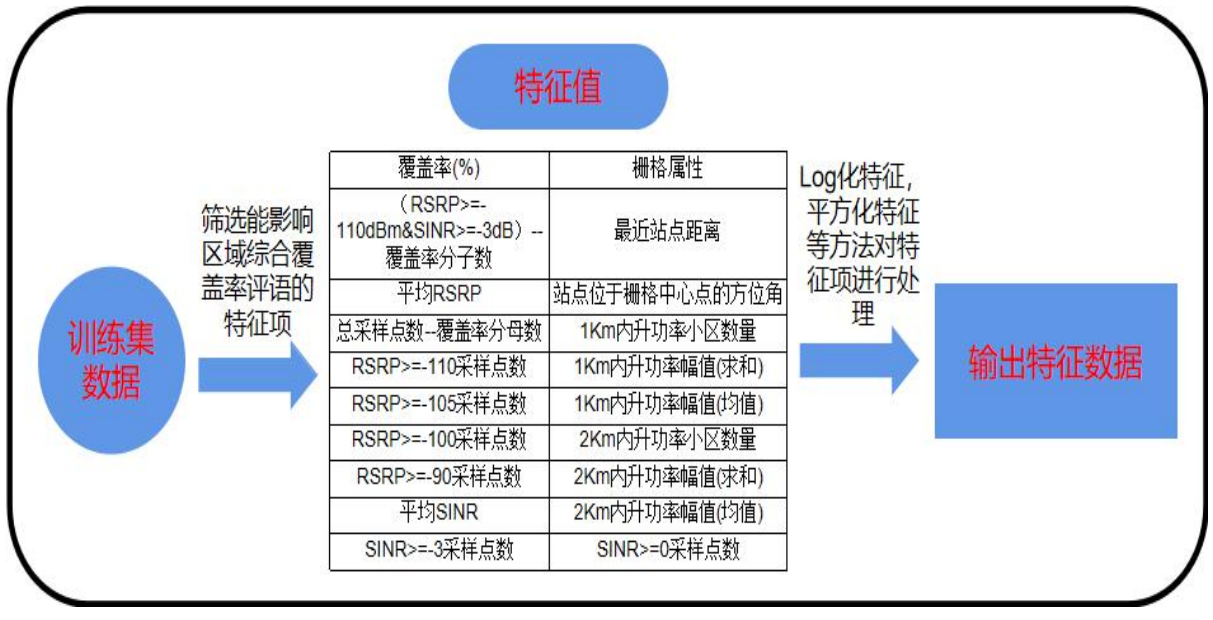
三、特征工程-数据处理

1.从前期已经完成功率优化的1453个FDD900区域摸底测试和复测的数据中，按70%/30%比例随机选取训练集和验证集，其中训练集数据用于机器学习特征值，验证集用于选择最佳的算法模型。并在训练集内用10折交叉验证做超参数调优:



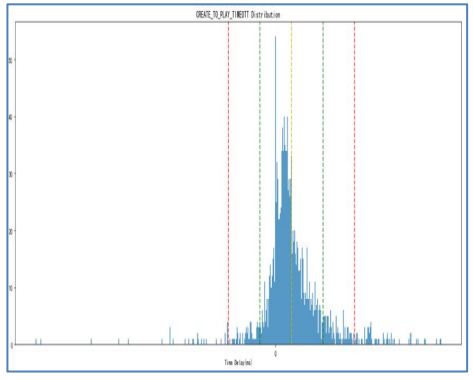
10折交叉验证

2.筛选能影响区域综合覆盖率评语的特征项，对栅格属性、覆盖率、平均RSRP等20项特征值进行包括Log化特征，平方化特征以及特征之间的加减多项式等多项处理：

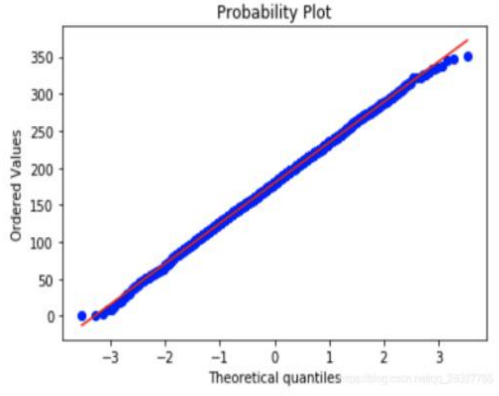


四、特征工程-异常处理与数据探索

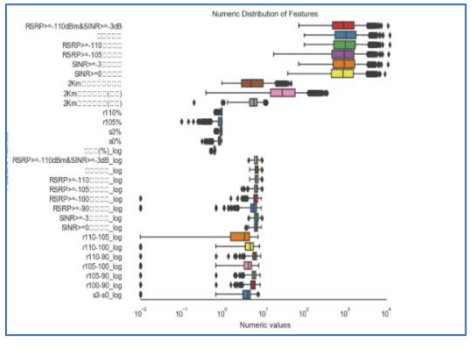
3.异常剔除：利用分布分析，箱型图分析，关联分析剔除训练集中强干扰的数据，并标准化处理：



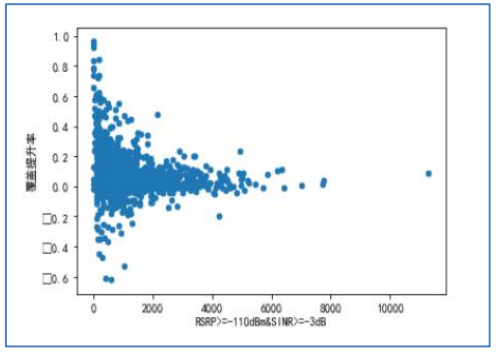
分布分析



QQ图分析正态性

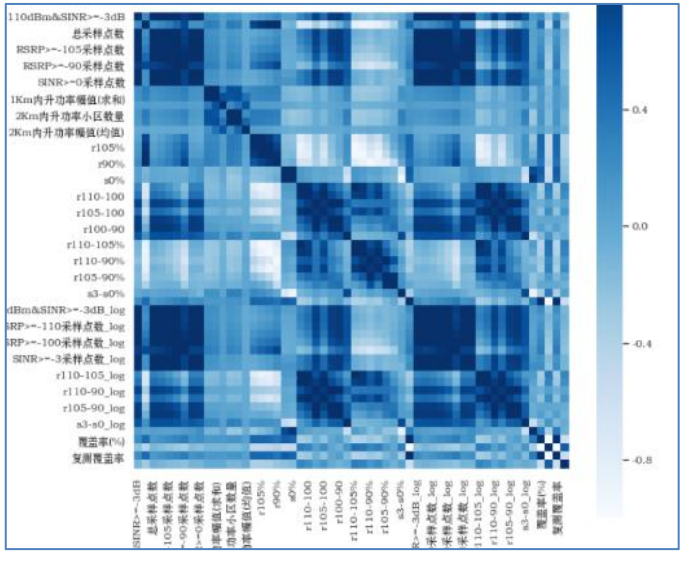


箱型图剔除异常值

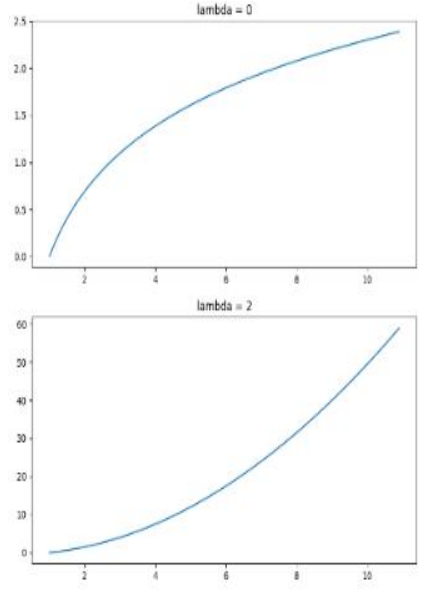


关联概率分析

4.利用box-cox转换技术，log1p公式对特征和标签做分布转换以减小泛化误差，并利用统计学pearson相关系数和spearman相关系数分析各个指标以及指标与覆盖提升率之间的关联度，剔除相关性比较大的特征，以降低对模型解释性和泛化性能的影响，并利用特征消除交叉验证做模型的特征筛选。



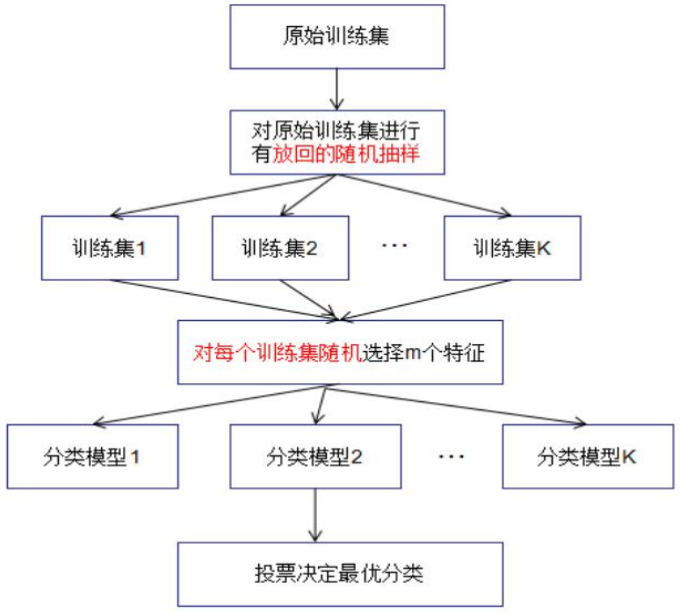
相关性分析



BOX-COX转换

五、算法设计

5.由于本课题数据规模在1万条以下，对比各种回归算法相关原理与使用场景，初步选取小数据规模下性能与准确度较好，且在比赛中大放光彩的的3种算法模型，包括：随机森林回归树算法，梯度提升回归树算法，岭回归算法。同时由于将覆盖提升率（取值范围为[-1,1]）作为标签数据，异常值对模型的性能影响较小，故并使用R Squared 作为模型评价系数：



随机森林原理

Algorithm 10.3 *Gradient Tree Boosting Algorithm.*

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

GBDT原理

$$\min_w ||Xw - y||_2^2 + \alpha ||w||_2^2$$

岭回归原理

$$R^2 = 1 - \frac{\sum_i \left(\hat{y}^{(i)} - y^{(i)}\right)^2}{\sum_i \left(\bar{y} - y^{(i)}\right)^2}$$

R Squared原理

六、成果算法优化与评价结论

6.利用网格搜索技术输出上述3个模型的学习曲线，并加入现网地市不断提交的复测数据以减少欠拟合。同时对随机森林和梯度提升决策树进行预剪枝和后剪枝，对岭回归进行正则化惩罚，并不断加入新的复测数据以降低过拟合，使用学习曲线挑选最优模型超参数。最后基于机器学习的结果，对三种模型算法进行优化，并输出结果

算法模型名称	算法结果验证
随机森林回归树算法模型	误差在5%以内的样本占比64%左右，误差在10%以内的样本占比85%左右。
梯度提升回归树算法模型	误差在5%以内的样本占比65%左右，误差在10%以内的样本占比86%左右。
岭回归算法模型	以0.15作为惩罚系数，验证集结果误差在5%以内的样本占比65.5%左右，误差在10%以内的样本占比90%左右。

各个算法对比

7.从左图算法的结果可知，结果值基本一致，但是考虑到奥卡姆剃刀原则（越简单的模型往往最后的泛化性能最好），**最后选取岭回归作为最终算法模型**。

岭回归-覆盖率差值(模拟评估-实际复测)分布--422个栅格		
差值区间	栅格数量	占比
1:-100%到-20%	0	0.00%
2:-20%到-10%	27	6.40%
3:-10%到-5%	67	15.88%
4:-5%到0%	190	45.02%
5:0%到5%	86	20.38%
6:5%到10%	25	5.92%
7:10%到20%	18	4.27%
8:20%到100%	9	2.13%

岭回归结果

谢 谢!

中国移动 China Mobile

5G++创新实训基地

技术实操的练兵场 · 能力认证的人才站 · 5G应用的孵化器