



中国研究生创新实践系列大赛  
“华为杯”第十六届中国研究生  
数学建模竞赛

学 校 西安交通大学

---

19106980112

参赛队号

---

1. 屈渝立

---

队员姓名 2. 王晓鹏

---

3. 张亚东

---

**中国研究生创新实践系列大赛**  
**“华为杯”第十六届中国研究生**  
**数学建模竞赛**

题 目      **无线智能传播模型**

---

**摘        要：**

随着 5G 时代的到来，5G 基站的部署已经逐渐展开。合理部署 5G 基站站址，对提高信号覆盖范围和信号质量有着重要的意义。无线传播模型作为无线网规划阶段的重要内容，对网络估算有着重要的作用。研究高效无线传播模型的建模方法对于正在开展的 5G 基站部署工作具有重要的实际工程意义。

本文以赛题提供的工程数据为驱动，首先进行数据可视化分析以及数据清洗，然后根据数学和物理模型进行低阶和高阶的特征提取，构建特征库，并且以此为基础，搭建 LightGBM 进行特征预筛选和模型训练，通过交叉检验等措施优化了模型参数，最终得到了一套完整的预测准确度高、泛化能力强的无线智能传播模型，并完成了该模型在华为云上的部署和正常测试，具有一定的实际应用价值。

问题一中，本文首先对测试集所有数据进行可视化，清洗了同一个坐标下重复的发射塔和测试点的数据，将数据拆分为发射塔和测试点两部分数据分别存储。然后根据从低阶到高阶、从局部到全局的思路展开特征工程，构建特征库。低阶特征主要参考了传统无线传播经验模型中的关键项；考虑到传统模型的缺陷，本文综合考虑地形环境和多个发射塔共同作用等因素的效果，进一步提取高阶特征。高阶特征分为局部和全局两个部分，考虑局部特征时，对于单个测试点来说，因为接收强度会受到附近的地形因素，例如对无线信号的遮挡、吸收、反射作用的直接影响，所以本文分析单个测试点周围的相对高度，并且借鉴二维卷积核的思想，将相对高度进行正负单独耦合以及相对高度的水平位置关系作为局部提取得到的特征。全局特征主要考虑到多个发射塔对测试点的共同作用，一方面本文按照距离从近到远找出最近的 M 个发射塔，将这些发射塔的基本特征和按照相对位置加权的发射功率进行平均化，另一方面按照以测试点为圆心，统计多个半径内的发射塔总数。最终得到相对完整的特征库，理论上共有 78 个特征供后续筛选。

问题二中，需要对问题一得到的特征进行筛选以及相关性排序，首先我们进行了特征计算的复杂度分析，发现高阶局部特征计算复杂度非常高  $O(X^3)$ ，( $X$  为测试点的总数)，考虑到比赛时间有限，只能舍弃该部分特征，实际计算得到的特征库共包含 68 个。对于相关性排序，本质上为输入特征对于预测结果的重要性程度，考虑到无线传播系统的复杂性，单纯计算向量的相关性无法很好的描述重要性程度，由于问题三中测试和部署得到的 LightGBM 表现和性能最为优良，同时基于提升树的方法具备从训练过程即树的生成和分裂中计算属性的重要性的优势，所以本文用 LightGBM 预训练后得到的重要性代表特征和目标之间的相关性。最终得到了所有特征的重要性，并提取出前 20 个最为重要的特征及重要性值，为问题三的模型进一步训练和调试做好准备。

问题三中，根据题目要求，本文构建一系列回归模型，例如多层感知机（MLP）、二维卷积神经网络（CNN）、XGBT、LightGBM，并且尽可能多的对各种方法进行性能和表现的比较和评估。测试比较后，为了兼顾计算效率和题目中指标的表现（RMSE, PCRR），本文采用了LightGBM算法来建立模型以实现RSRP的预测。通过合理的参数配置和调试，在线下的大范围验证集（5%，也就是25万个数据点）进行检验，得到的最终RPRS均方根误差（RMSE）在5.3左右波动，弱信号覆盖率（PCRR）在0.45至0.6之间，满足题目要求。

# 目录

1 问题重述 .....	4
1.1 问题背景 .....	4
1.2 需要解决的问题 .....	5
2 模型假设 .....	6
3 问题一的分析与建模 .....	7
3.1 数据可视化及预处理 .....	7
3.2 特征工程 .....	8
3.2.1 低阶特征 .....	8
3.2.2 高阶特征 .....	9
4 问题二的分析与建模 .....	14
5 问题三的分析与建模 .....	17
5.1 机器学习算法及对比 .....	17
5.1.1 经典回归系列 .....	17
5.1.2 卷积神经网络 .....	17
5.1.3 多层感知机模型（MLP） .....	18
5.1.4 提升树系列 .....	18
5.2 模型参数调节与设置 .....	20
5.3 模型结果 .....	20
6 模型评价及展望 .....	23
参考文献 .....	24

# 1 问题重述

## 1.1 问题背景

5G 技术的不断推进，使得全球逐渐进入 5G 网络部署热潮，图 1.1 给出了未来十年中国 5G 用户规模预测图。合理选择基站站址，以提高覆盖区域并确保信号强度是运营商部署 5G 网络时的重要任务。合理选择基站站址作为无线网络规划中的重要内容，依赖于通过有效的无线传播模型产生的网络估算结果。无线传播模型是指通过对目标通信覆盖区域内的无线电波传播特征进行预测，以估算小区覆盖范围、小区间网络干扰以及通信速率等。建立有效且精确的无线传播模型是极其复杂的，这是因为在实际工程中，无线电传播环境复杂，会受到传播路径上各种因素的影响，如平原、山体、建筑物、湖泊等，使得电磁波产生复杂的透射、绕射、散射、反射等，传播方式和传播路径也变得不再单一。

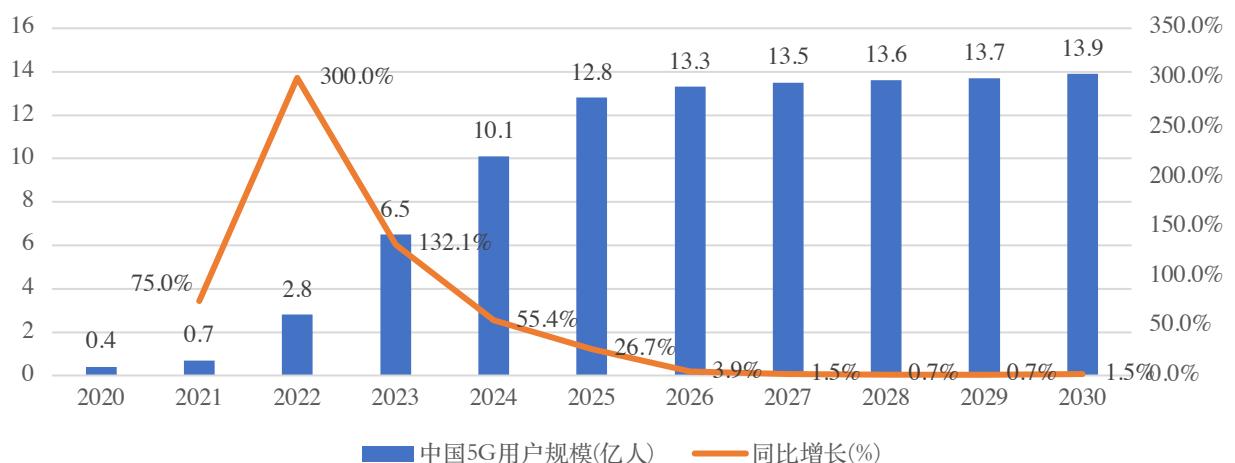


图 1.1 2020-2030 年中国 5G 用户规模预测

(数据来源: <http://www.elecfans.com/d/855567.html>)

现有的无线传播模型通常分为经验模型、理论模型和改进型经验模型。经验模型的获得是从经验数据中获取固定的拟合公式，典型的模型有 Cost 231-Hata、Okumura 等。理论模型是根据电磁波传播理论，考虑电磁波在空间中的反射、绕射、折射等来进行损耗计算，比较有代表性的是 Volcano 模型。改进型经验模型是通过在拟合公式中引入更多的参数从而可以为更细的分类场景提供计算模型，典型的有 Standard Propagation Model (SPM)。

改进型经验模型是当前的主流建模方法，但是该方法需要对传播场景进行划分，并且每个场景对应一个相对固定的公式进行拟合，很难直接适用于 5G 网络中。在实际传播模型建模中，为了获得符合目标地区实际环境的传播模型，需要收集大量额外的实测数据、工程参数以及电子地图用来对传播模型进行校正。

近年来，随着大数据、人工智能、机器学习的兴起，以神经网络为代表的智能算法由于其强大的泛化能力以及从复杂系统中提取有效信息的能力，其在各个领域出色的表现越来越受到人们广泛的关注，例如智能家居、人脸识别、自动驾驶等等。“数据”也逐渐因为其量级的提升以及智能算法的普及而为人类带来意想不到的价值和收获。实际环境中的无线传播本身是一个极其复杂的问题，受到发射源、接收点以及环境的各种因素影响，甚至可以说整体是一个非常复杂的动力学系统。而对复杂系统来说，机器学习恰恰具备优越的分析和泛化能力，同时具备非线性度高、实时性强、容易工业化使用等优势，也就是说，基于大数据的 AI 模型为当今时代解决智能无线传播提供了巨大的可能性，其背后的工业价值和实际价值更是不可估量。

## 1.2 需要解决的问题

问题一：高效的机器学习模型建立依赖于输入变量与问题目标的强相关性，因此输入变量也称为“特征”。特征工程的本质是从原始数据中转换得到能够最好表征目标问题的参数，并使得各个参数的动态范围在一个相对稳定的范围内，从而提高机器学习模型训练的效率。根据公式（1）中的 Cost 231-Hata 模型以及数据集信息设计合适的特征，并阐述原因。

表 1.1 数据的字段含义

字段名称	含义	单位
Cell Building Height	小区站点所在栅格(Cell X, Cell Y)的建筑物高度，若该栅格没有建筑物，则为 0	m
Cell Altitude	小区站点所在栅格(Cell X, Cell Y)的海拔高度	m
Cell Clutter Index	小区站点所在栅格(Cell X, Cell Y)的地物类型索引	-
Cell X	小区所属站点的栅格位置，X 坐标	-
Cell Y	小区所属站点的栅格位置，Y 坐标	-
Height	小区发射机相对地面的高度	m
Azimuth	小区发射机水平方向角	Deg
Electrical Downtilt	小区发射机垂直电下倾角	Deg
Mechanical Downtilt	小区发射机垂直机械下倾角	Deg
Frequency Band	小区发射机中心频率	MHz
RS Power	小区发射机发射功率	dBm
Cell Index	小区唯一标识	-
X	栅格位置，X 坐标	-
Y	栅格位置，Y 坐标	-
Building Height	栅格(X,Y)上的建筑物高度，若该栅格没有建筑物，则为 0	m
Altitude	栅格(X,Y)上的海拔高度	m
Clutter Index	栅格(X,Y)上的地物类型索引	-
RSRP	栅格(X, Y)的平均信号接收功率，标签列	dBm

问题二：基于提供的各小区数据集，设计多个合适的特征，计算这些特征与目标的相关性，并将结果量化、排序，形成表格，并阐明设计这些特征的原因和用于排序的量化数值的计算方法。特征名称及其与目标的相关性。

问题三：在设计和选择了有效的特征之后，根据建立的特征集以训练数据集，建立基于 AI 的无线传播模型来对不同地理位置的 RSRP 进行预测。

## 2 模型假设

考虑到实际情况及处理的可行性，我们做出如下假设：

1. 假设提供的数据是真实可靠的。
2. 假设数据中的异常点（数据标记错误等）是少量的，且对结果不产生重大影响。
3. 假设如果栅格点建筑物高度不为 0，则测量地点为建筑物楼顶。
4. 假设所有数据（训练集和测试集）是根据统一的参考点定位得到的栅格点坐标。

### 3 问题一的分析与建模

根据题目，实际环境中的无线传播本身是一个极其复杂的问题，受到发射源、接收点以及环境的各种因素影响，甚至可以说整体是一个非常复杂的动力学系统。对于给定测试点 RSRP 的预测，传统模型过度依赖经验公式，其众多经验参数在不同的地理因素、环境因素等情况下，需要大量的调试和归纳，耗费人力物力的同时由于模型非线性度往往较低导致估算实际平均信号接收功率 (Reference Signal Received Power, RSRP) 精度无法满足更高的要求。另外，通常经验公式中只包含一个发射塔，无法很好的适应实际中，测试点受到多个发射塔影响的情况。

与传统经验公式不同，基于大数据的机器学习具备对复杂系统优越的分析和泛化能力，同时有着非线性度高、实时性强、容易工业化使用等优势。对于问题一，本文通过对无线传播的大数据进行深度挖掘，通过“数据预处理——特征提取”构建为后续的选取和模型训练的所需特征库。

#### 3.1 数据可视化及预处理

本文将赛题提供的训练集和参考测试集文件混合，如图 3.1，作出测试点地物类型在二维平面的分布图，并且作出局部的测试点和发射塔的分布图，如图 3.2。



图 3.1 训练集所有测试点的地物类型分布图

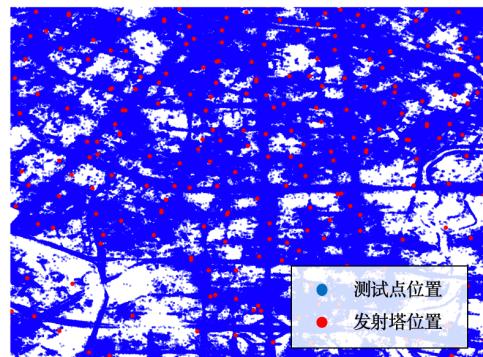


图 3.2 局部测试点和发射塔分布图

可以看到，虽然不同发射塔和测试点可能来自不同的数据文件，但是总体呈现较为密集的全局分布。进一步的，分别以发射塔和测试点的坐标为唯一 ID 进行分析并且绘制重

复频次统计直方图（图 3.3 和图 3.4），可以看到存在大量的坐标重合，那么需要对重合的样本进行清洗。对于同一个坐标点出现多个发射塔数据，可以看到发射天线的高度和功率存在差异，除了发射塔所在的地物编号之外，对发射塔的所有相关参数进行平均化处理。同时，对于同一个出现的多个 RSRP 数据，考虑到可能是人为因素和环境影响，所以对重复的 RSRP 求取平均值作为该点的最终 RSRP。

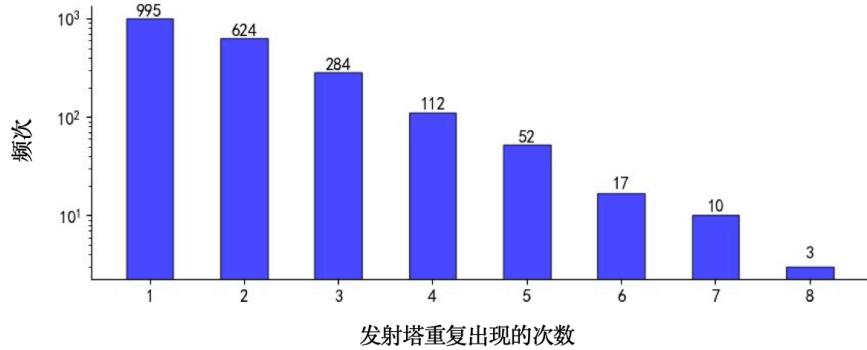


图 3.3 发射塔坐标重复出现的次数频数直方图

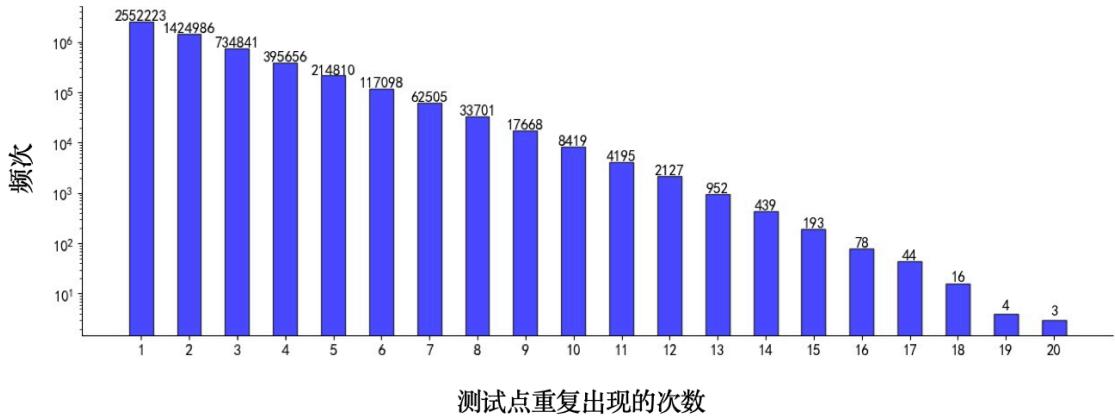


图 3.4 测试点坐标重复出现的次数频数直方图

通过以上的分析可以看出，原始数据中的测试点和发射塔并不存在严格的一一对应关系，或者说，具体某个测试点在全局的分布图当中，会受到多个发射塔的影响，所以我们将来原始数据的每一行剥离成为发射塔数据和测试点数据，清洗后保证了两类数据中不存在重复点，进行分别存储，大大削减了重复数据对存储空间的占用，以及为后续机器学习模型的训练和测试做最基本的准备。

## 3.2 特征工程

如何根据已知数据以及问题背景构造并且选取合适的特征是“大数据+机器学习”模型的核心问题。本小节用数据预处理后的发射塔和测试点两类数据，分析“低阶到高阶、局部到全局”的实际情况并且搭建合理的数学模型，进行特征库的构建。

### 3.2.1 低阶特征

低阶特征参考了传统无线传输损耗模型，将经验公式当中用到的单独项提取成为特征，例如距离等，以及根据测试点所在栅格的地物索引，进行分类并且转为 one-hot 编码，具体

分析和实现如下：

### (1) 根据传统公式提取简单特征

对传统的无线传输损耗模型进行分析，例如 Okumura-Hata、COST 231-Hata、Egli Path Loss、Standard Propagation Model，可以看到对于发射塔和接收点，经验公式大多与有效高度有关，并且对距离和高度等参数进行 log 变换，所以相对应的，我们将这部分从经验公式提取的单独项作为特征。具体计算公式如下：

$$H_t = A_{cell} + B_{cell} + H_{cell} \quad (3-1)$$

其中， $H_{cell}$ ——发射塔的有效高度（单位：m，不做特殊说明，距离单位均为米）， $A_{cell}$ ——发射塔所在海拔高度， $B_{cell}$ ——发射塔所在位置的建筑物高度， $H_{cell}$ ——发射天线的自身到发射塔底部的高度。

$$H_r = A_r + B_r \quad (3-2)$$

其中， $H_r$ ——测试点的有效高度， $A_r$ ——测试点所在海拔高度， $B_r$ ——发射塔所在位置的建筑物高度。

根据经验公式，我们对带宽、距离相关的参数进行对数变换，此处不一一列出。

### (2) 地物类型归纳

不同的地物类型会对电磁波的传播产生不同的干扰，考虑到题目给出的 20 个地物类型具有一定的属性重叠，例如，海洋、内陆湖泊、湿地这三类地物都属于含水量丰富的类型，所以把他们归纳为“水域”这一新的属性，并且以此类推，归纳得到下表所示类型，并且进行 0-1 编码。

表 3.1 归纳后的地物类型 0-1 编码表

地物类型	水域	城市开阔区域	植被区域	建筑集中区
海洋，内陆湖泊，湿地	1	0	0	0
城郊开阔区域，市区开阔区域，道路开阔区域	0	1	0	0
植被区，灌木植被，森林植被	0	0	1	0
城区超高层建筑，城区高层建筑，城区中高层建筑，城区高密度建筑群	0	0	0	1
其他	0	0	0	0

## 3.2.2 高阶特征

对于此问题来说，考虑到无线传播的特性，每个测试点会受到局部地形建筑物的影响，建筑和房屋越密集、周围建筑越高，对于信号的遮挡和反射效果越严重；另外，测试点的 RSRP 会受临近发射塔的影响，临近发射塔发射功率越高、发射塔越密集，那么测试点的信号强度往往更高。考虑到上述两点特性，我们从局部地理环境因素和全局发射塔因素两个层面，分别建立数学模型进行特征提取：

### 3.2.2.1 局部地理环境因素——在测量点附近楼房建筑的属性

#### (1) 局部相对高度耦合

测试点附近的局部环境主要和建筑物高度以及海拔高度有关，为了简化计算，我们以测试点为中心，找出附近方形区域内的所有已知临近点的有效高度（即  $H_r$ ），没有高度数据的点用 0 表示，如图 3.5 (a) 所示，以测试点（黑色）为方形区域中心，找出相邻一个

栅格的所有其他测试点的有效高度即蓝色栅格（称为已知栅格），没有测试过的点用灰色栅格（称之为未知栅格）表示。然后计算临近有效栅格与中心点的相对高度如图 3.5 (b) 所示，橙色栅格表示该点比中心测试点的建筑高，绿色表示该点比中心低，根据常识可以知道，橙色栅格对于黑色测试点的信号强度呈现的遮挡及反射作用比较强，绿色栅格的对于中心的影响呈现的效果与橙色完全不同，所以我们将其划分成两个部分，如图 3.5 (c)、(d) 所示，分别计算 c 和 d 的平均值，得到两个新的特征。

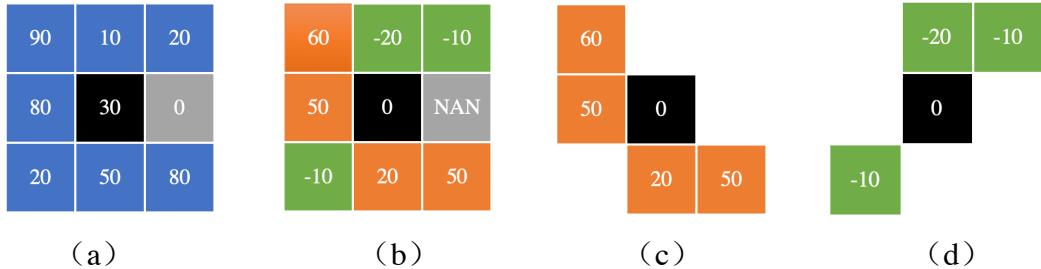


图 3.5 局部高度特征分析示意图

具体数学模型如下：

1	2	3
4	5	6
7	8	9

图 3.6 局部测量点方阵下标示意图

假设中心测试点的有效高度为  $h_{center}$ ，所提取的局部特征的仿真栅格半径为  $r$ ，图 3.6 即为  $r = 1$  的情况，栅格的索引从左上角开始标号为  $\{1, 2, 3, \dots, (2r+1)^2\}$ ，对应的有效高度为  $\{h_1, h_2, h_3, \dots, h_{\frac{(2r+1)^2+1}{2}}, \dots, h_{(2r+1)^2}\}$ ，其中  $h_{center} = h_{\frac{(2r+1)^2+1}{2}}$ ，那么可以得到相对高度  $\Delta h_i = h_i - h_{center}$ ，然后根据  $\Delta h_i$  的正负得到两个下标集合：

$$PosHSet = \{i \mid \Delta h_i \geq 0, i \in \{1, 2, 3, \dots, (2r+1)^2\}\} \quad (3-3)$$

$$NegHSet = \{i \mid \Delta h_i < 0, i \in \{1, 2, 3, \dots, (2r+1)^2\}\} \quad (3-4)$$

进而计算均值：

$$PosH = \frac{1}{PosNum} \sum \Delta h_i, i \in PosSet \quad (3-5)$$

$$NegH = \frac{1}{NegNum} \sum \Delta h_i, i \in NegSet \quad (3-6)$$

得到的  $PosH$  和  $NegH$  分别表示临近测试点局部区域内的建筑物等效正负相对高度的平均值，用于衡量局部范围内的环境遮挡和反射效果。该指标本质上借鉴了二维卷积的核心思想，该数学模型通过耦合局部的特征，得到小范围内地理环境的对信号传播的评估特征。

## （2）局部相对高度一维化

考虑到小节（1）计算的  $PosH$  和  $NegH$  无法较好地描述相对位置关系，而中心测试点可能由于附近发射塔和周围地形的耦合作用，导致相同的  $PosH$  和  $NegH$  值却呈现不同的信号传播效果，如图 3.7 所示。

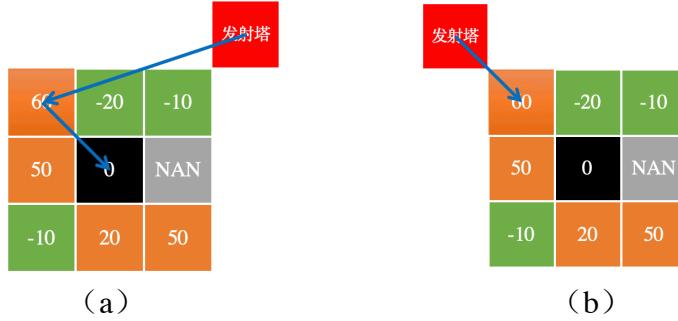


图 3.7 发射塔对相同的局部高度的不同作用效果

图 3.7 (a) 所示, 相对高度为 60 的建筑对于发射塔的无线信号主要呈现反射作用, 而图 3.7 (b) 中, 该建筑却对于发射塔的无线波主要呈现遮挡作用, 也就是说局部的相对高度的相对位置也需要考虑, 我们对相对高度矩阵进行一维向量化作为特征的一部分, 如图 3.8 所示。

1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---

图 3.8 相对高度方阵一维向量化的下标 (对应图 3.6)

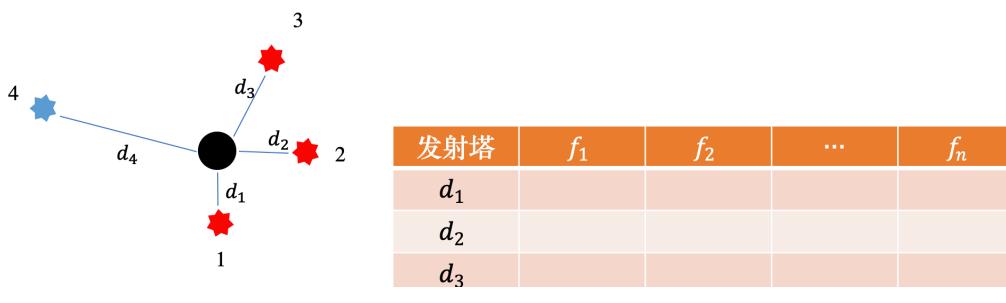
另一方面, 从图 3.7 可以看出, 不同位置的发射塔对测试点的影响受到附近局部地理环境的影响可能会呈现截然不同的效果, 所以在接下来我们对全局特征进行提取。

### 3.2.2.2 全局发射塔对测试点的影响

从上面的分析了解到受到环境因素、传播特性影响, 实际中一个测试点的 RSRP 会受到全局发射塔的影响, 为了简化问题, 突出重要因素, 我们分析最接近于目标测试点的 M 个发射塔的特征。

#### (1) M-Nearest 发射塔特征

从前面的数据预处理和可视化可以看出, 每个测试点可能会受到附近多个发射塔的影响, 考虑到理论上, 发射塔和测试点之间距离越远, 收到的信号越弱, 我们根据距离找出最邻近该测试点的 M 个发射塔, 并且对发射塔的各组参数进行汇总, 如图 3.9 (a) 所示, 黑色圆形为测试点, 星形为发射塔, 取 M=3 为例, 找出最邻近该测试点的 3 个发射塔 (红色), 然后按照距离从近到远列出发射塔相关所有特征, 如图 3.9 (b) 所示。可以知道, 每个测试点都对应了一组最近邻发射塔的特征矩阵, 我们将其一维向量化, 作为 M-Nearest 发射塔特征的基础部分。



(a) 分布示意图

(b) 数据格式

图 3.9 单个测试点的 M 最邻近发射塔基础特征

$$DistanceSet = \{[n_1, n_2, n_3, \dots, n_i, \dots, n_N] \mid d_{n_{i-1}} \leq d_{n_i} \leq d_{n_{i+1}}\} \quad (3-7)$$

其中，N——发射塔的总数，n——发射塔的索引下标， $d_{n_i}$ ——发射塔 $n_i$ 到测试点的距离。得到的 $DistanceSet$ 集合为按照从近到远的发射塔新的索引，我们取出前 M 个进行基础特征的提取。

在基础特征的前提下，进一步分析发射塔在测试点的叠加功率及合成角度。如图 3.10 所示，示意图中取 M=3，具体数学模型如下。

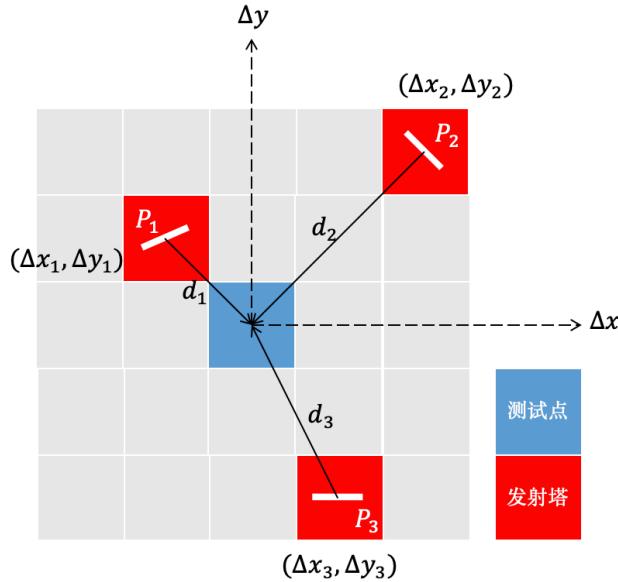


图 3.10 单个测试点的 M 最邻近发射塔疊加示意图

假设以该测试点为参考点，M 个发射塔的相对坐标为  $\{(\Delta x_i, \Delta y_i) | i \in \{1, 2, 3, \dots, M\}\}$ ，计算疊加效果：

$$(Q_x, Q_y) = \sum \frac{1}{P_i} (\Delta x_i, \Delta y_i) \quad i \in \{1, 2, 3, \dots, M\} \quad (3-8)$$

并且进行极坐标转化：

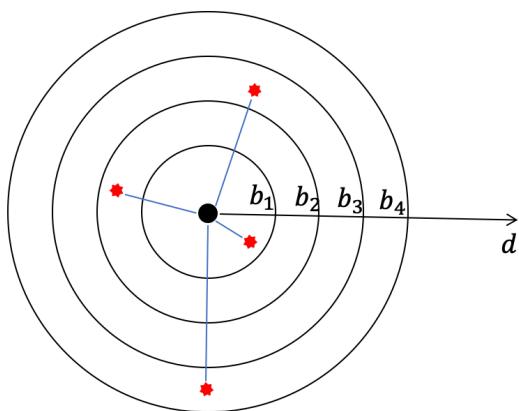
$$Q = \sqrt{Q_x^2 + Q_y^2} \quad (3-9)$$

$$\theta_q = \arctan \frac{Q_y}{Q_x} \quad (3-10)$$

这两个指标用于描述最邻近测试点的发射塔在测试点的功率强度疊加以及合成偏向角。

## (2) 以测试点为圆心的发射塔环形个数分布

前文描述了最近邻发射塔的特征，与此同时需要考虑以测试点为圆心，如图 3.11 (a) 所示，各个圆环范围内的发射塔个数。



(a) 圆环示意图

距离 $d$ 范围	个数
$d < b_1$	
$d < b_2$	
$d < b_3$	
$d < b_4$	

(b) 数据结构

图 3.11 单个测试点的 M 圆环范围内的发射塔统计个数

## 4 问题二的分析与建模

通过对问题一的分析，我们可以列出所有特征，并且计算特征的复杂度，如表 4.2 所示，表中底色为淡黄色的特征都进行了对数变换，X 为测试点的总数，Y 为发射站的总数，根据数据统计分析可知， $X \gg Y$ 。由复杂度分析可知，计算 PosH 和 NegH 即以测试点为中心的仿真区域平均正相对高度和负相对高度的复杂度非常高，由于时间限制，无法在比赛期间内较快的计算获取该两个特征，所以暂时舍弃 PosH 和 NegH。

根据机器学习模型 LightGBM 训练可以得到前 20 个与结果 RSRP 预测相关性最强的特征，以及重要程度，如表 4.1 所示。

表 4.1 相关性最强的前 20 个特征名称及相关性值

排序	特征名称	相关性
1	nearest_theta	4563
2	X	3840
3	nearest_distance_0	3642
4	Y	3381
5	nearest_distance_1	3368
6	nearest_distance_2	3159
7	nearest_Q	2991
8	nearest_distance_4	2912
9	nearest_distance_3	2640
10	neighbor_num_5000	2516
11	neighbor_num_2000	2348
12	neighbor_num_4000	2184
13	Azimuth_0	2179
14	neighbor_num_3000	2123
15	Altitude	1897
16	Azimuth_4	1843
17	Azimuth_1	1807
18	Height_0	1755
19	Azimuth_3	1668
20	Azimuth_2	1659

表 4.2 所有特征名称、含义解释及复杂度

名称	含义解释	复杂度
X	测试点横坐标	-
Y	测试点纵坐标	-
Altitude	测试点海拔	$O(X)$
Building Height	测试点的建筑物高度	$O(X)$
Clutter Index	地物索引	-
Effective_Height	测试点有效高度	$O(X)$
Azimuth_0	距离该测试点最近的发射塔水平偏向角	$O(X*Y)$
Azimuth_1	距离该测试点第二近的发射塔水平偏向角	$O(X*Y)$
Azimuth_2	距离该测试点第三近的发射塔水平偏向角	$O(X*Y)$

Azimuth_3	距离该测试点第四近的发射塔水平偏向角	O(X*Y)
Azimuth_4	距离该测试点第五近的发射塔水平偏向角	O(X*Y)
Height_0	距离该测试点最近的发射塔高度	O(X*Y)
Height_1	距离该测试点第二近的发射塔高度	O(X*Y)
Height_2	距离该测试点第三近的发射塔高度	O(X*Y)
Height_3	距离该测试点第四近的发射塔高度	O(X*Y)
Height_4	距离该测试点第五近的发射塔高度	O(X*Y)
Electrical Downtilt_0	距离该测试点最近的发射塔垂直电倾角	O(X*Y)
Electrical Downtilt_1	距离该测试点第二近的发射塔垂直电倾角	O(X*Y)
Electrical Downtilt_2	距离该测试点第三近的发射塔垂直电倾角	O(X*Y)
Electrical Downtilt_3	距离该测试点第四近的发射塔垂直电倾角	O(X*Y)
Electrical Downtilt_4	距离该测试点第五近的发射塔垂直电倾角	O(X*Y)
Mechanical Downtilt_0	距离该测试点最近的发射塔垂直机械倾角	O(X*Y)
Mechanical Downtilt_1	距离该测试点第二近的发射塔垂直机械倾角	O(X*Y)
Mechanical Downtilt_2	距离该测试点第三近的发射塔垂直机械倾角	O(X*Y)
Mechanical Downtilt_3	距离该测试点第四近的发射塔垂直机械倾角	O(X*Y)
Mechanical Downtilt_4	距离该测试点第五近的发射塔垂直机械倾角	O(X*Y)
Frequency Band_0	距离该测试点最近的发射塔带宽	O(X*Y)
Frequency Band_1	距离该测试点第二近的发射塔带宽	O(X*Y)
Frequency Band_2	距离该测试点第三近的发射塔带宽	O(X*Y)
Frequency Band_3	距离该测试点第四近的发射塔带宽	O(X*Y)
Frequency Band_4	距离该测试点第五近的发射塔带宽	O(X*Y)
RS Power_0	距离该测试点最近的发射塔发射功率	O(X*Y)
RS Power_1	距离该测试点第二近的发射塔发射功率	O(X*Y)
RS Power_2	距离该测试点第三近的发射塔发射功率	O(X*Y)
RS Power_3	距离该测试点第四近的发射塔发射功率	O(X*Y)
RS Power_4	距离该测试点第五近的发射塔发射功率	O(X*Y)
Cell Altitude_0	距离该测试点最近的发射塔海拔高度	O(X*Y)
Cell Altitude_1	距离该测试点第二近的发射塔海拔高度	O(X*Y)
Cell Altitude_2	距离该测试点第三近的发射塔海拔高度	O(X*Y)
Cell Altitude_3	距离该测试点第四近的发射塔海拔高度	O(X*Y)
Cell Altitude_4	距离该测试点第五近的发射塔海拔高度	O(X*Y)
Cell Building Height_0	距离该测试点最近的发射塔所在建筑物高度	O(X*Y)
Cell Building Height_1	距离该测试点第二近的发射塔所在建筑物高度	O(X*Y)
Cell Building Height_2	距离该测试点第三近的发射塔所在建筑物高度	O(X*Y)
Cell Building Height_3	距离该测试点第四近的发射塔所在建筑物高度	O(X*Y)
Cell Building Height_4	距离该测试点第五近的发射塔所在建筑物高度	O(X*Y)
Effective_Height_0	距离该测试点最近的发射塔有效高度	O(X*Y)
Effective_Height_1	距离该测试点第二近的发射塔有效高度	O(X*Y)
Effective_Height_2	距离该测试点第三近的发射塔有效高度	O(X*Y)
Effective_Height_3	距离该测试点第四近的发射塔有效高度	O(X*Y)
Effective_Height_4	距离该测试点第五近的发射塔有效高度	O(X*Y)
nearest_distance_0	到该测试点最近的发射塔的距离	O(X*Y)
nearest_distance_1	到该测试点第二近的发射塔的距离	O(X*Y)

nearest_distance_2	到该测试点最三近的发射塔的距离	$O(X^*Y)$
nearest_distance_3	到该测试点最四近的发射塔的距离	$O(X^*Y)$
nearest_distance_4	到该测试点最五近的发射塔的距离	$O(X^*Y)$
neighbor_num_500	到该测试点距离 500 以内的发射塔的个数	$O(X^*Y)$
neighbor_num_1000	到该测试点距离 1000 以内的发射塔的个数	$O(X^*Y)$
neighbor_num_2000	到该测试点距离 2000 以内的发射塔的个数	$O(X^*Y)$
neighbor_num_3000	到该测试点距离 3000 以内的发射塔的个数	$O(X^*Y)$
neighbor_num_4000	到该测试点距离 4000 以内的发射塔的个数	$O(X^*Y)$
neighbor_num_5000	到该测试点距离 5000 以内的发射塔的个数	$O(X^*Y)$
nearest_Q	距离该测试点最临近 5 个发射塔功率的加权平均	$O(X^*Y)$
nearest_theta	距离该测试点最临近 5 个发射塔角度的加权平均	$O(X^*Y)$
water_index	是否为水域，是则为 1	$O(X)$
broaden_index	是否为宽阔区域，是则为 1	$O(X)$
plant_index	是否为植被区，是则为 1	$O(X)$
bulding_index	是否为建筑密集区，是则为 1	$O(X)$
RegH	以测试点为中心的方阵区域平均正相对高度	$O(X^*X^*X)$
PosH	以测试点为中心的方阵区域平均负相对高度	$O(X^*X^*X)$
deltaH_1	以测试点为中心的方阵区域 1 号相对高度	$O(X^*X^*X)$
deltaH_2	以测试点为中心的方阵区域 2 号相对高度	$O(X^*X^*X)$
deltaH_3	以测试点为中心的方阵区域 3 号相对高度	$O(X^*X^*X)$
deltaH_4	以测试点为中心的方阵区域 4 号相对高度	$O(X^*X^*X)$
deltaH_6	以测试点为中心的方阵区域 6 号相对高度	$O(X^*X^*X)$
deltaH_7	以测试点为中心的方阵区域 7 号相对高度	$O(X^*X^*X)$
deltaH_8	以测试点为中心的方阵区域 8 号相对高度	$O(X^*X^*X)$
deltaH_9	以测试点为中心的方阵区域 9 号相对高度	$O(X^*X^*X)$

## 5 问题三的分析与建模

无线电传播模型的建模问题本质上是一个回归问题，即平均信号接收功率的回归问题。但是由于无线电传播过程受到多种因素的影响，基于前述工作提取了多维的数据特征，利用传统方法对特征进行建模并不能取得理想的效果，而机器学习作为解决高维特征问题的重要工作，是该问题的重要解决思路。

机器学习包含传统的机器学习（例如决策树、支持向量机等）以及深度学习，深度学习在计算机视觉与语音处理等领域取得了重要成就，而传统的机器学习在工业大数据问题中具有不错的表现。本文对多种模型进行了有益的尝试，并最终选定 Lightgbm (Light Gradient Boosting Machine)作为主要模型。

### 5.1 机器学习算法及对比

虽然数据和特征将会对最终的回归效果产生巨大的影响，但机器学习算法的合理设计也是尤为重要的。经过长期的发展，如今的机器学习算法种类繁多，对于本问题的回归任务而言，可以利用传统机器学习中经典的 logistic regressive, Ridge Regression, Lasso Regression 回归算法等，也可以利用数据科学领域大放异彩的 Boosting 系列，例如 GBDT、XGBoost、LightGBM 等，也可以利用多层次感知机与卷积神经网络等神经网络模型。本文对上述三种类型中的某一种或某几种进行了测试，并最终选择 Lightgbm 作为模型生成算法。

#### 5.1.1 经典回归系列

本文对 logistic regressive 算法进行了测试，logistic regressive 是传统的多项式回归算法，建立多项式模型，然而回归多项式需要自行设计，会消耗很多的时间和人力成本，考虑到时间限制及鲁棒性等因素，最终未选为模型生成算法，本文仅对其模型特性进行简单描述。

Logistic regressive 模型在训练过程中训练效率较高，但其效果较差，表现为相较于后文中的模型，均方根误差 (RMSE) 很大为 14.53，且弱覆盖识别率小于 5%，不满足问题要求。通过对预测结果的分析，模型更加倾向于拟合值集中在测试集均值附近。

#### 5.1.2 卷积神经网络

在测试阶段，卷积神经网络未利用前文所提取特征，而是将每个小区的数据当成一张多通道的“图片”，利用二维卷积神经网络进行预测。神经网络设计为“图”到“图”的神经网络，即输入与输出拥有相同形状。网络设计描述如下：

输入层为 3 通道 2000\*2000 三维矩阵，矩阵生成方式为：发射塔位置为图片中心位置，其他位置按照与发射塔的相对位置进行映射，各个通道分别为测试点的建筑物高度，测试点的地物类型，测试点的海拔高度。其中，样本中未涉及的点所有参数均为 0。

隐含层为 3\*3 卷积残差神经元，即利用 (1\*1), (3\*3), (1\*1) 三种类型的卷积核组成的残差神经单元，经过残差神经单元后，矩阵大小不改变，仅改变通道数。隐含层共采用了 3 个残差神经单元，每个神经单元的输出通道数分别为 64, 32, 1。激活函数为 ReLU 激活函数。

输出层为全连接层，即将最后一个神经元的输出直接展开，作为输出。

训练前，对所有数据进行归一化处理，网络训练的参数为：优化器选择 sgd (随机梯度下降) 优化器，迭代次数设置为 100 次，学习率设置为 0.05，损失函数为 mse (mean square error 均方根误差)，当最优结果保持次数超过 5 次后训练自动停止。训练集选择两千个小

区，每次输入两个小区数据进行计算，网络在含有两块 1080Ti GPU 26G 显存的服务器上运行，本次训练共持续 31 次，因为网络误差不再下降而停止，每次迭代时间为 11 分钟左右。最终 RMSE(Root Mean Square Error 均方根误差)在测试集上为 10.45，RSRP 为 0.13。

考虑到卷积神经网络的训练时间过长，且比赛时间限制，我们最终放弃选择卷积神经网络。在后续工作中虽然发现新的数据集分割方法，即通过对全局数据集处理进行再分割，但是由于时间限制，未能进行进一步实现。但是我们认为卷积神经网络是有效的，因此在特征提取阶段，本文运用了卷积的思路提取特征。

### 5.1.3 多层感知机模型（MLP）

MLP 作为反向传播网络在多特征拟合中也有重要应用，因其理论上可以对任何函数进行拟合而受到广泛关注。在算法选择阶段，本文建立了简单的多层感知机模型，模型描述如下：

输入层为 17 个神经元，对应于原始的 17 个特征。（后改为 20 个神经元，对应于选择的 20 个特征）

隐含层含有 3 个全连接层，神经元个数分别为 32,64,8，激活函数为 sigmoid 函数。

输出层为 1 个神经单元。

训练前，对所有数据进行归一化处理，网络训练的参数为：优化器选择 sgd (随机梯度下降) 优化器，迭代次数设置为 1000 次，学习率设置为 0.01，损失函数为 mse (mean square error 均方根误差)。训练集选择 100 个小区数据点，网络在 1050Ti GPU 2G 显存电脑运行，网络迭代次数为 1000 次，迭代时间为 10min，RMSE 为 9.42，RSRP 为 0.21。（17 个原始特征阶段结果）

多层感知机的结果相对于前述方法有所提高，但是多层感知机的神经元及层数设计缺乏理论指导，其参数调节过程较为复杂。

### 5.1.4 提升树系列

GBDT (Gradient Boosting Decision Tree)是在决策树基础上通过迭代构建起来的梯度提升树算法，其独特的算法设计，使其具有极为优秀的拟合能力和泛化能力，在回归任务中具有突出表现。XGBoost 是一个用 C++ 开发的提升树框架，属于 boosting 集成学习里面的一种，由陈天奇等人提出，该框架继承了 GBDT 等算法那，并对其运行过程进行了优化，提高了运算速度与运行效率，成为重要工业机器学习算法及人工智能比赛中的重要方法。本文利用 XGBoost 中集成的 GBDT 算法作为算法选择阶段的测试方法，取得了包括上述三种算法在内的四种算法中的最优结果。另外相较于前三种方法，XGBoost 需要调整的参数较少，更加适用于当前场景。

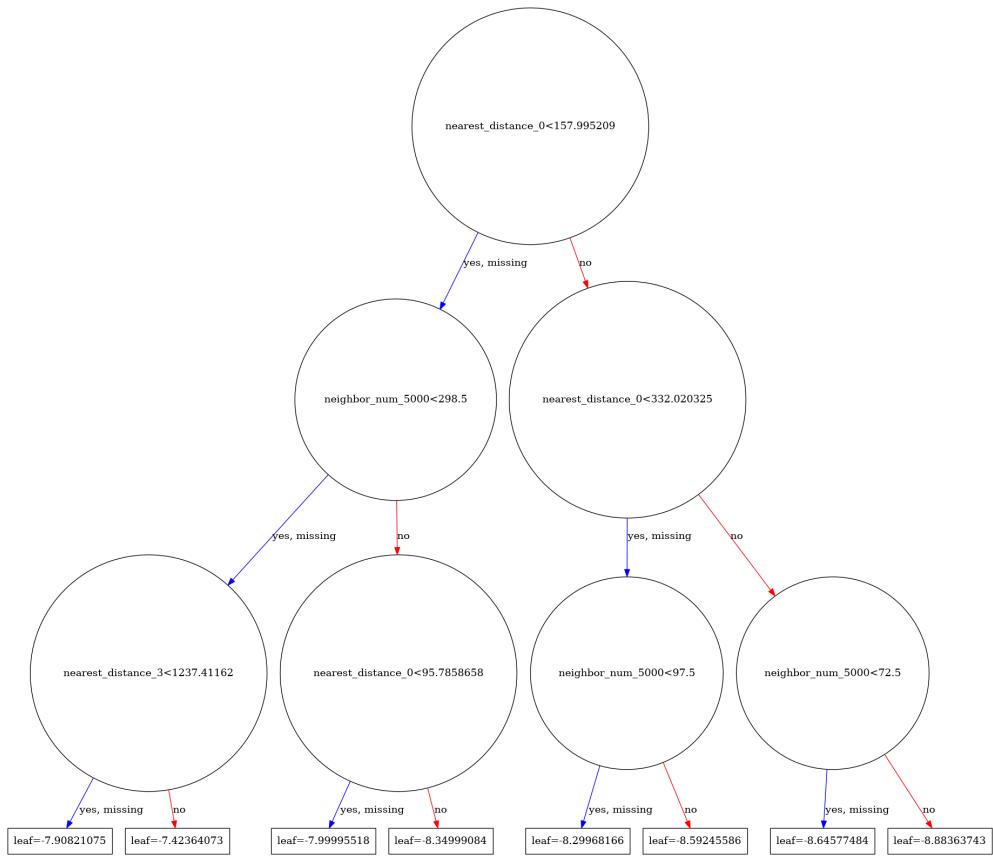


图 5.1 XGBoost 部分生成树结构

XGBoost 在前期算法选择过程中上产生了较好的结果（100 个小区数据结果，17 个原始特征），其平均信号接收功率的 RMSE 在测试集上为 RMSE 为 8.94，RSRP 为 0.25，训练时间为 4 分钟左右。在后续工作中，我们发现 XGBoost 在大数据集上训练速度较慢，若利用全部数据进行训练，预估需要耗费近两日时间。通过后续研究，选择 LightGBM(Light Gradient Boosting Machine)作为 XGBoost 的改进框架，该框架可以在不损失准确率的前提下有效降低训练时间，其在全部数据上的训练时间仅为 2 小时左右。我们最终选择 LightGBM 作为树分类模型的建模方法。由于 XGBoost 的建模方式与参数调节方式与 LightGBM 大致相同，此处不再过多叙述，在后文中将介绍 LightGBM 的具体建模过程。

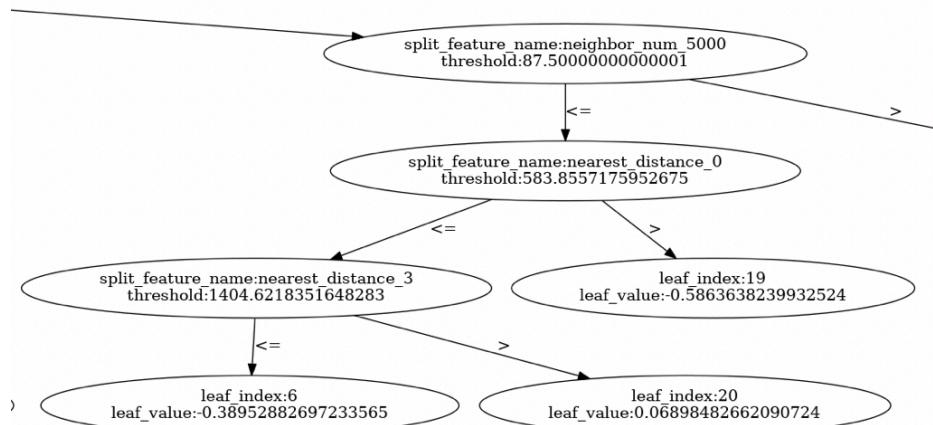


图 5.2 LightGBM 部分生成树结构

## 5.2 模型参数调节与设置

模型的输入参数为 20 个特征。LightGBM 参数如表 5.1 所示，对于可调解参数，我们采用基于贪心的调参方法——坐标下降法，即每次直言这一个维度搜索，当将该参数调节到最优时，我们继续调节其他参数。根据经验及对时间因素的考虑，我们对参数 max\_depth（调节范围为[4-10]步长为 0.1）、eta（调节范围为[0.01-0.19]步长为 0.3）、gamma([0-0.5]步长为 0.1)、alpha ([0-0.5]步长为 0.1)、其他参数取典型值。最终确定的参数为：max\_depth 为 7，eta 为 0.1，gamma 为 0.2，alpha 为 0.2。

表 5.1 LightGBM 参数含义

参数名	含义	典型值
eta	学习率	0.01~0.2
min_child_weight	一个子集的所有观察值的最小权重和	1
max_depth	树的最大深度，值越大，树越大，模型越复杂	3~10
gamma	分裂节点时，损失函数减小值只有大于等于 gamma 节点才分裂，gamma 值越大，算法越保守，越不容易过拟合。	0
alpha	L1 正则化，增加该值会让模型更加收敛	0
objective	需要被最小化的损失函数	reg:linear
subsample	构建每棵树对样本的采样率	1
colsample_bytree	列采样率，也就是特征采样率。	1
colsample_bylevel	构建每一层时，列采样率。	1

图 5.3 给出了利用最优参数进行训练时的 RMSE 下降曲线。

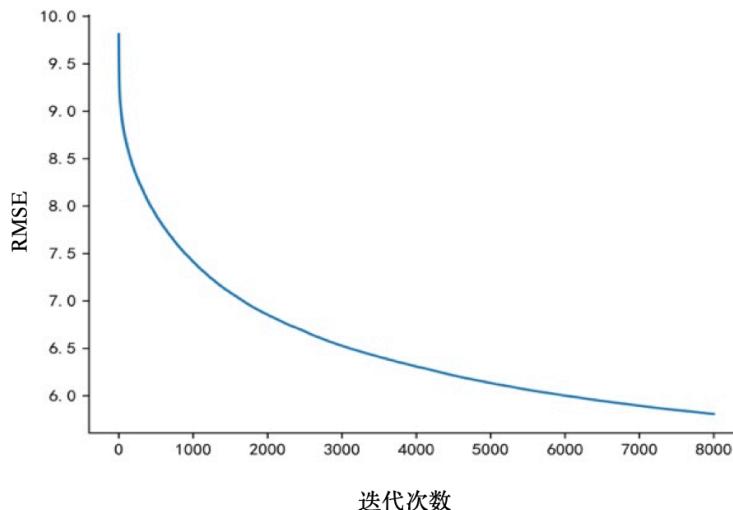


图 5.3 LightGBM 在验证集上的 RMSE 下降曲线

## 5.3 模型结果

按照上述参数，我们得到了模型文件，下面以文件“train\_set/train\_124701.csv”中小区数据为例，对模型结果进行展示。

本文建立模型在该测试中的表现优异，各项数值指标如表 5.2 所示，为了便于直观的表示预测结果，我们给出了该测试集上的 RSRP 的预测结果及实测结果的热力图，如图 5.4 所示。按照数据集中的数据顺序，我们给出了如图 5.5 所示的预测结果与实测结果的比较。

表 5.2 该小区的评价指标值

指标名称	Recall	Precision	PCRR	RMSE
值	0.3670	0.3960	0.3810	7.0395

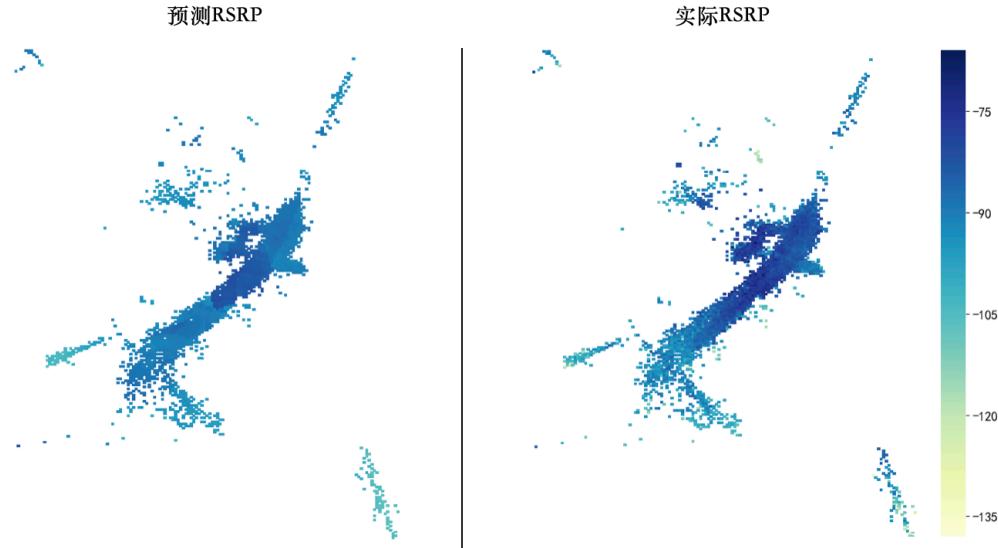


图 5.4 测试集 RSRP 预测结果及实际结果热力图

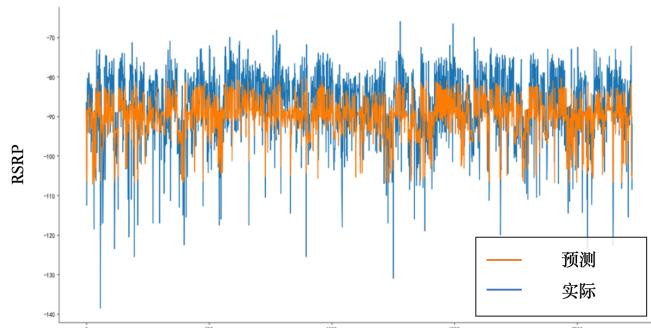


图 5.5 预测值和实际值的分布图

从数值指标及图 5.4 我们可以看出，模型的 RMSE 较好，预测结果表现良好，且拥有较高的弱覆盖率，通过图 5.4 可以发现，预测结果的整体效果较好。图 5.5 表征预测结果的波动范围相对实测结果较小，这可能是由于模型在当前训练情况下仍然处于欠拟合状态，可以进一步提升，另外，相对较小的波动在全局上可能具有更好的泛化能力。

表 5.3 不同模型在同一验证集的表现对比

	Recall	Precision	PCRR	RMSE
LightGBM	0.5556	0.7339	0.6324	5.182
XGBoost	0.1597	0.6765	0.2584	7.302
MLP	0.1397	0.5432	0.2223	7.951

为了进一步说明我们建立模型的结果，我们将 LightGBM 与 XGBoost 与 MLP 进行对比。通过表 5.3 可知，LightGBM 的各项参数均优于其他两项，说明本文建立的模型是更有

效且准确的。

## 6 模型评价及展望

本问题本质上属于数据挖掘，在这类问题中，数据，特征工程和算法三者缺一不可。而我们这次分别就这三个方面对问题进行了分析，特别是在特征工程阶段，创新性的通过建立多个数学模型，提出了许多有效的高阶特征，并且通过选取最合适的算法在最终的任务上取得了十分显著的效果，优点主要表现在：

1. 细致，合理的数据预处理；
2. 与实际问题紧密结合的特征提取和选择；
3. 创新性的建立多个数学模型来进行特征工程；
4. 准确且高效的算法模型选择；
5. 模型泛化能力强，测试表现稳定。

对于本问题而言，本文认为接下来工作重心仍然是放在特征工程上面，特别是针对高阶特征的选取往往不是简单的组合能够得来，可以深入利用数学工具和物理模型，进一步研究本问题背后隐含的数理逻辑关系，由于问题本身的复杂性，仅仅靠经验公式来刻画背后的规律是不太可行的，所以机器学习仍然是解决这类问题的最有效的方式，但是建立精准高效的特征模型确是我们可以做到的。同时，算法的选择和优化也是需要重点关注的，要同时兼顾准确率和效率，在算力足够的情况下，深度学习更是适合解决这类问题的一大工具海量的数据。

## 参考文献

- [1] Popoola S I , Atayero A A , Faruk N , et al. Standard Propagation Model Tuning for Path Loss Predictions in Built-Up Environments[J]. 2017.
- [2] Laiho J , Wacker A , Tomáš Novosad. Radio Network Planning and Optimisation for UMTS[M]. John Wiley & Sons, Inc. 2001.
- [3] 丁俊民, 廖振松. 基于大数据挖掘的 4G 网络规划研究[J]. 信息通信, 2016(2):246-247.
- [4] 宋媛媛,王萍, 张庆芳,等.基于最小二乘法的 TD-LTE 传播模型校正研究[J]. 电子测量技术, 2015, 38(1):123-125.
- [5] 于仰源, 孙宜军, 王磊, et al. 一种基于 MR 数据修正无线传播模型的方法[J]. 移动通信, 2019(3).
- [6] 钱小康, 何华忠. 3GPP 3D 无线传播模型在 5G 基站覆盖预测中的应用[J]. 上海信息化, 2018(11):69-72.