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	复旦大学，浙江工商大学
参赛队号	19102460049
队员姓名	1.刘展宏
	2.陆恒
	3.胡佳琪

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 **关于无线智能传播模型的研究**

摘 要：

在传统无线网络规划过程中，受限于数据搜索条件，对于无线网络的规划结果往往不尽如人意，而当下无线网络传播环境的复杂性较以往也是大大增加。但随着科学技术的发展，近年来通过 LTE 大数据以及大数据 AI 机器学习等已经成为了助力研究无线信道传播的重要手段之一。本文主要研究的是通过传统无线传播模型以及现有的大量数据，通过机器学习技术训练使得得到平均信号接收率（RSRP）有较小均方误差，为有效建设提供帮助的同时也能够达到提高传播效率的目的。

针对任务一，首先我们需要了解相关传输背景，基于这些背景以及对传统传播模型的了解，利用传统模型以及传统模型中的校正因子，并且结合已知工程参数，地图数据并获取通过已知数据可以求得的衍生数据，计算数据形成一个完整体系。并将之归纳分类为五种特征，分别为：**已有特征，衍生特征，计算特征，校正特征以及涉及特征**。对于每一部分特征都进行详细解释和阐述，并展现了其计算公式，通过理论分析以及运用 cost31-hata 等模型对部分特征进行计算并进行合理性检验，从完整性和方向性上构建**整体特征体系**。

针对任务二，基于任务一中的特征体系，我们对其进行是否可以通过已有数据进行计算筛选得到可达特征，同时出于主客观性整体把握对地物类型，cost231-hata-RSRP,相对高度进行相关性分开判别。其余特征再通过发散性分析以及相关性分析的手段对特征数据进行量化排序。在相关性分析中，共运用了三种特征系数计算方式，将相关性进行量化排序。宏观把握特征的整体表现，微观细化特征的具体相关性。横向比较不同系数计算方法对特征排序的影响，纵向比较特征排序内含的不变性。因此达到了对不同特征的相关性进行**多层面分析**的要求，为后续机器学习做好了铺垫。

针对任务三，出于选择最优模型考虑，我们采用了考虑了单个模型以及混合模型对数据的影响，单个模型中选取了**全连接神经网络**，**随机森林**模型线上线下进行 RSRP 的均方误差 RMSE 的计算，混合模型则考虑了**深度神经网络联合线性回归模型**，再次进行线上线下的测试。其中神经网络相关模型线下选择 200 万作为训练集，100 万作为测试集；随机森林线下选择 1100 万数据作为训练集，所有模型线上均选择 1200 万作为测试集，具体的 RMSE 值如下所示，其中联合模型，因其在线下的表现已不如线下全连接神经网络而随机森林表现与全连接神经网络表现差异不明显，因此并未对其进行线上得分评估。

模型	全连接神经网络	深度神经网络联合线性回归	随机森林
线下 RSRP 值的 RMSE	9.698	9.953	9.584
线上 RSRP 值的 RMSE	9.434	/	/

关键词：整体特征体系 多层面分析 全连接神经网络 联合模型 随机森林

一．问题重述

1.1 问题背景

随着无线网络技术不断进步和发展，人们对于无线网络的需求不断增加的同时用户对于网络的依赖性更是在不断的加深，这现象的发生无疑是用户对通信服务提出更高要求的一种最为直接的表象化呈现方式。

无线通信技术现在正在经历着它有史以来最为快速的发展时期，而要研究移动通信系统工程的首要工作就是要明白无线传播其本身的特性，其次是要研究无线传播信道所带来的相关特性，基于对这些特性的了解才能够更好的在建立城市通信网络时进行规划。根据3GPP（3rd Generation Partnership Project,第三代合作计划）组织的最新计划和安排[1]，预计2020年，5G将会迎来全球化的商业部署，中国则大概率会提前实施这一计划。作为5G网络部署的重要一环，无线网络的规划问题即对5G基站的覆盖研究是基站的具体部署计划的基础，因此对于无线传播模型也提出了更高要求[1],传播模型的准确度将会直接影响到无线规划的准确度包括基站规模估算，链路预算等等。通过相关研究可以发现，无线通讯系统主要受移动传播环境的影响与制约，而传播环境自然是复杂多变的，理论上在不同地区，不同城市内的不同地物地貌，地类等，每个区域都映射了不同的传播模型，可以称之为“指纹”模型，因为这种特殊性导致无法精准的给出一个广泛适用模型，传统的传播模型也只是利用研究方法的不同进行分类，分为经验模型，理论模型，改进经验模型，经典模型有很多例如Cost 231-Hata模型，Okumura模型等等，他们通过经验数据来获取固定的拟合公式，从而完成对路径损耗的计算，作为传播模型的一种评估指标。而近年来，伴随着大数据技术的发展，我们期望，可以运用相关历史数据，通过AI机器学习技术，将其运用于无线网络当中去,完成对传播路径损失的计算，实现在不同场景下预测平均信号接收率（RSRP），从而提高了网络传输效率的同时又节省了建设成本。

1.2 待解决问题

➢**问题一**：要求我们根据已知的传统经验模型Cost 231-Hata以及提供的相关数据的具体信息，设计特征并阐述理由

➢**问题二**：需要我们通过多个层面判断上述特征设计的合理性，并将这些有效且合适的特征通过一定的量化方法进行排序并阐述理由。

➢**问题三**：要求我们运用这些有效特征，通过AI机器学习技术对不同地理位置的RSRP建立合适的预测模型并详述建模过程与方法。

二．问题分析

2.1 问题一

考虑到最终我们需要用机器学习技术，为了得到更好的机器学习的效率，如何将输入变量与我们的目标之间的相关性提到最高是我们最终能否得到较好的AI传播模型的关键，因此，我们在设计这些有一定相关性的变量，即特征时，需要有一个整体的特征体系来参

考。即，我们搭建的整体的特征体系是作为整个传播模型的基础，因此，对整个特征体系提出了更高的要求，需要尽可能的满足一定完整性，需要尽可能的包含可能会对模型产生影响的各种因素，需要利用已知数据通过数学手段等来得到未知数据等等。

基于在对传统传播模型有一定了解的基础上，考虑到传统模型中涉及到的参数对路径损失的影响，同时也需要考虑到不同传播环境下模型校正中所运用的校正因子带来的影响。同时，通过原有数据集的数据属性所作出的衍生变换也可能是影响最终结果特征中的一部分，当然数据本身也是不可忽略的重要组成部分。 上述部分都是以数据或是模型本身为出发点，若作反向设计，那么根据模型和数据可以得到 RSRP 值和 PL 值并将之与实际值进行比较，之间的差值本身也是一种对 RSRP 表现的反应。前者以理论为依据出发，后者以实践为依据出发，从理论和实践中共同探索构建整体特征体系，在一定程度上达到了我们对于完整性和多方向性的要求，为后续工作打下夯实基础。

2.2 问题二

基于上述整体的特征体系，我们的主观性判断错误或是过多的特征冗余亦或是特征与目标之间相关性不高，特征数据过于离散等等都会造成最终机器学习效率的低下，因此特征的提取和构建最终特征平台就显得尤其重要。

首先我们以问题一为基础，从中挑选可以通过数据进行计算的特征，成为可达特征，对这部分可达特征的数据进行数据过滤，填补等预处理，将处理好的特征数据先进行离散化分析再进行相关性分析，考虑到数据本身属性，所属类别，量纲等不同所带来的影响，我们考虑可以运用三种相关性系数进行量化排序的手段：皮尔逊相关性系数，斯皮尔曼相关性系数，肯德尔相关性系数，全方面的对数据的相关性表现进行分析，分析其中的特征数据表现的原因，从而选择合适的特征作为重要特征，标记为后续机器学习的重要指标。

2.3 问题三

基于问题二中所提及的特征，就单个模型而言，我们拟采用全连接神经网络模型，并配套使用最为合适的优化器，激活函数，选择合适的节点，层数等进行线上线下对于 RMSP 均方误差的计算。除此之外，我们还使用随机森林模型进行计算。就多模型联合而言，我们拟采用深度学习线性回归器，即将全连接网络与线性回归相融合再次进行线上线下均方差计算。基于线上 1200 万训练集，以及线下 200 万训练集，100 万测试集计算的均方误差下进行 RMSE 计算后，后者模型线下与前者模型中线下出现的最小值进行对比，若小于前者则进行线上计算，若大于或等于则不考虑线上计算，从而完成在最有效时间下选择最优的模型的目标。

三．模型假设

1.假设建筑物上方放置的发射天线具有高辐射中心线
2.在电磁波的传播过程中没有其他人为因素进行干扰
3.数据中并不包含当时季节，气候等对于传播造成的影响

四．符号说明

符号	含义
PL	路径损耗
h_b	建筑物高度
h_s	基站离地高度
h_g	基站海拔高度
h_m	移动天线离地高度
h_{mg}	移动天线海拔高度
h_r	移动天线有效高度
h_t	基站天线有效高度
f	频率
d	收发天线之间的水平距离（链路距离）
w	街道宽度
L_F	自由空间路径损耗
$A_{mu}(f,d)$	自由空间中值损耗
$G(h_t)$	基站天线有效高度增益因子
$G(h_r)$	移动天线有效高度增益因子
G_{AREA}	环境增益
$a(h_r)$	接收天线有效高度修正因子
K_{mr}	郊区校正因子
Q_o	开阔地区校正因子
C_m	场景校正因子
K_{street}	街道修正因子
Q_r	准开阔地区校正因子
R_u	农村校正因子
K_h	丘陵地校正因子
K_{sp}	一般倾斜地形校正因子
Δh	表示的是地形起伏高度差
h_{min}	表示的是地形起伏的最小高度
K_{im}	孤立山峰校正因子
$Diff$	绕射损耗
υ	绕射常数
$L_{clutter}$	地物损耗

五．问题 1:信道传播工程中的特征设计

通过研究相关文献[2]，我们发现无线电波在不同环境中的传播特性是传播模型研究及其优化的首要问题，而无线电波信号在传播过程中出现的损耗决定了无线传输能够到达的最远距离，换句话说，预测无线电波的路径损耗问题是研究其覆盖率的首要目标，该损耗值在一定程度上也体现了较大范围内接受信号功率的平均值的变化趋势。因此，我们以路径损失为目标，需要探究可能与路径损失有关系的各个数据特征。我们主要考虑从两方面入手，一则是已有的传统传播模型及其改进模型等，在模型中涉及到的参数或是校正因子都可能作为最后计算损失时的有效特征，二则是根据已有的数据信息，通过对于电波传播的理解以及一定的三维立体计算，可以利用已有数据得到衍生特征，在对这些衍生特征进行计算时，得到的计算值也可以更加清晰化我们的计算过程，也更有利于我们研究其内在规律。

5.1 常见传播模型

5.1.1 传播模型研究背景

要搭建整体特征体系就需要明白并了解并明白传播损耗的具体含义，传播损耗主要分为：平滑地面，不规则地形传播损耗；绕射损耗；穿透损耗，反射损耗。绕射损耗的计算较为复杂，随着不同的绕射常数而变化，而穿透损耗则通常与室内所处位置，具体阻隔物等有关，反射损耗则是与电与磁接触的地面性质有关。现有的传播模型一般是运用于预测地形，障碍物，认为环境对电磁波路径损耗的影响，因此基于现有的传播模型，我们主要对室外传播损耗进行研究。首先我们通过查阅文献[2]发现在自由空间下的传播损耗表现，即当两个天线处在自由空间下时，其路径损耗计算如下所示[6]:

$$PL = 32.4 + 20logf + 20logd$$

路径损耗用PL表示，其余参数如下:

f:频率 d:收发天线之间的水平距离

自由空间即是电磁波指各向同性，无吸收，且处于电导率为 0 的均匀介质中。

只有有了自由空间下的损耗模型，我们才可以考虑在不理想的或者是平坦状态下的电磁波传播情况。与自由空间损耗相比，平坦地面传播路径损耗计算如下:

$$PL = 40logd - 20logh_s - 20logh_m$$

h_s:基站天线离地高度 h_m:移动天线离地高度

但在现实生活中，绝对的平坦地形是不可能存在的，需要根据传播地形，传播环境等进行更细化的分类。就传播环境而言，已知[2]，在电波传播预测模型中，主要可以分为两类：室内传播模型和室外传播模型。在实际传播环境中，室外传播模型又可以分为两种：宏蜂窝模型以及微蜂窝模型。出于对传播环境多样性的考虑，以下简单对三类环境进行简单介绍：宏小区，微小区，微微小区。

宏小区：宏蜂窝，大区制蜂窝，指面积很大的区域，覆盖半径大约在 1-30km,基站发射天线通常设置在周围建筑物上方。收发之间没有直达射线。路径损耗主要由移动台附近的平屋顶的绕射或散射决定。

微小区：微蜂窝，覆盖半径大约在 0.1-1km,发射天线的高度一般和周围建筑物相同，略高于或低于。损耗来源主要由周围建筑物的绕射和散射，类似于在周围建筑物形成的峡谷内进行传播。

微微小区：典型尺寸在 0.01-0.1km。发射天线一般在屋顶下面或者建筑物内。可以发现主要是指室内或者室外的极小 的范围之内进行传播。

现在城市中，运用最为广泛的即是宏蜂窝环境下的传播预测模型，基于宏蜂窝理论中

发射天线具有高辐射中心线理论假设成立的情况下，我们发现电波主要由两种部分组成，一种是沿屋顶垂直向下的垂直平面波，一种是沿街道水平芳香的水平平面波。并且在基站附近，以垂直平面波为主，在远离基站的城市环境中则以水平面波为主。

通过阅读相关文献，有关于城市宏蜂窝地区传播路径的损耗模型的通用函数如下：

$$PL = f(h_b, h_{BS}, h_m, w, f, d)$$

各参数如下所示：

h_b：建筑物高度 w：街道宽度 h_t：基站天线有效高度 h_r：移动天线有效高度

我们注意到上述通用函数中用到的是基站有效高度，接收天线有效高度等，因此需要对有效高度进行表示与计算，一般来说，可以通过几何运算表示有效高度，具体示意图如下：

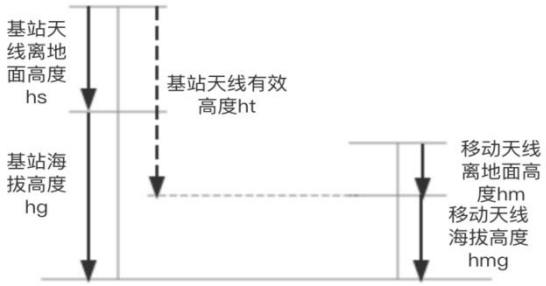

图 1.基站天线有效高度示意图

若用公式表示，即如下：

$$h_t = h_s + h_g - h_{mg}$$

其中，各参数意义如下：

h_g：基站海拔高度 h_{mg}：移动天线海拔高度

5.1.2 常见的传播模型及其校正

现有的传播模型主要分成两种，一种为确定性模型，一种为经验模型。确定性模型是需要对现场环境运用电磁理论的科学计算方法，即基于几种射线跟踪的电磁理论的精确方法对路径损耗进行预测，显然这将无线网络规划问题转化为了一种及其复杂的电磁问题，因为其高度的数学复杂程度，我们在现实中更多的采用的是经验模型。所谓经验模型，就是利用大量的结果统计分析后导出的数学拟合公式，是基于在某种条件下的数学运算。

以下，主要介绍几种常见的经验模型

①Okumura 模型（奥村模型）：

以下简称奥村模型，奥村模型是预测城区信号是最为广泛使用的模型，应用频率在 150MHz-1920MHz 之间，可扩展至 3000MHz，距离为 1km-100km 之间，天线高度在 30m-1000m 之间[3]。模型主要公式如下所示：

$$PL = L_F + A_{mu}(f, d) - G(h_t) - G(h_r) - G_{AREA}$$

L_F：自由空间路径损耗　$A_{mu}(f, d)$：自由空间中值损耗

$G(h_t)$：基站天线有效高度增益因子$G(h_r)$：移动天线有效高度增益因子G_{AREA}：环境增益

其中，关于天线高度增益计算如下：

$$G(h_t) = 20log(\frac{h_t}{200}), 30m < h_t < 1000m$$

$$G(h_r) = \begin{cases} 10log(\frac{h_r}{3}) & h_r \leq 3m \\ 20log(\frac{h_r}{3}) & 3m < h_r < 10m \end{cases}$$

说明天线高度增益是关于基站天线高度的严格的增函数,且处在不同移动天线高度下，增益的效果亦是不同。

➢**Okumura-Hata 模型**（校正后的奥村经验模型）：

出于奥村模型是奥村等人在东京近郊用宽范围的频率及几种固定的基站天线高度，移动高度下形成的一系列曲线，基于大量实测数据进行拟合得到的经验模型，是以准平坦地形大城区的中值场强为参考，所以若要运用在中国还需要对该模型进行参数校正，通过最小二乘法和线性回归的方法，得到校正后的经验模型如下：

市区：

$$PL = 69.55 + 26.16lgf - a(h_r) - 13.82lgh_t + (44.9 - 6.55lgh_t)lgd$$

$a(h_r)$:接收天线有效高度修正因子，不同城市规模下该修正因子的表现不尽相同，具体表现如下：

表 1.市区 Okumura-Hata 模型下有效高度修正因子表达式

城市规模		接收天线有效高度修正因子
大城市	$f > 300MHz$	$a(h_r) = 3.2(lg11.75h_r)^2 - 4.97$
	$f < 300MHz$	$a(h_r) = 8.29(lg1.54h_r)^2 - 1.1$
中小型城市		$a(h_r) = (1.1lgf - 0.7)h_r - (1.56lgf - 0.8)$

值得注意的是，在运用该模型时，做出了三点假设：

1.两个全向天线之间的传播损耗

2.作为准平滑地形而不是不规则地形

3.一城市市区的传播损耗为标准，其他地区采用校正公式

因此，我们得到其他地区下的传播损耗公式如下：

郊区：

$$PL(countryside) = PL - 5.4 - 2[lg(f/28)]^2$$

PL 是指在市区的损失值，为加以区分，我们把$-5.4 - 2(lgf/28)^2$称作郊区校正因子K_{mr}用PL(countryside)来表示郊区下的路径损失值

开阔地区：

$$PL(wide) = PL - 40.98 - 4.78[lg(f)]^2 + 18.33lgf$$

同理，我们把$-4.78(lgf)^2 + 18.33lgf - 40.94$称之为开阔地区校正因子$Q_o$，此处用 PL(wide)表示开阔地区的路径损失值。研究表明，校正后的模型下计算得到的损失值与仿真数据大致相同，因此，我们认为该校正模型足够可靠，且另一方面表明，有效高度校正因子对最终的路径损失值的测量起着非常关键的作用。

②Cost 231-Hata 模型：

在对上述模型描述的过程中，我们发现，上述不论奥村模型或是其改进模型都不适用于覆盖距离小于 1km 的个人通信系统，因此在计算中上述模型存在着预测值偏高的问题，

而 Cost 231-Hata 模型可以解决这个问题，为了实现对 1800MHz-1900MHz 频段的无线信号覆盖，Cost 231-Hata 模型将频率增大到 2000MHz,适用于大城市中的高密度区[[4]。

Cost 231-Hata 模型是在 Hata 模型的基础上进行一定程度的扩展得到的，同时，通过阅读相关文献，我们发现该模型具体使用范围如下：频率在 1500MHz-2000MHz 之间，适用于小区半径大于 1km 的宏蜂窝系统，发射天线有效高度在 30m-200m 之间，接收天线高度在 1-10m 之间。具体损耗经验公式如下：

$$PL = 46.3 + 33.9lgf - 13.82lgh_t - a(h_r) + (44.9 - 6.55h_t)lgd + C_m$$

C_m:场景校正因子

虽然在上述公式中，与奥村模型中运用到了相同参数例如移动天线有效高度，但是需要注意的是，有效高度的计算在不同传播模型下是不同的，而有效高度的计算也有多种方法，例如：在基站周围 5-10 公里的范围内可以去地面海拔高度的平均，或者地面海拔高度的地形拟合线等，这些取决于传播模型以及我们要求的计算精度。

在已知不同传播模型下不同有效高度的不同计算方法的条件下，因此有了针对于有效高度在该 cost231-hata 模型下的校正因子的表达，关于接收天线有效高度校正因子在不同规模城市下的具体表现如下：

表 2.cost231-hata 模型下接收天线有效高度修正因子表达式

城市规模	接收天线有效高度修正因子
大城市	$a(h_r) = 3.2(lg11.75h_r)^2 - 4.79$
中小型城市	$a(h_r) = (1.11lgf - 0.7)h_r - (1.56lgf - 0.8)$
$h_r = 1.5m$	0

关于场景校正因子的表现如下：

表 3.场景校正因子

类别	大城市中心	城市	郊区	农村
校正因子值	3	0	-12.28	-22.52

比较上述两种模型，发现[5],两者最大的区别在于两者有不同的衰减系数，后者的频率衰减系数为 33.9，前者衰减系数为 24.16,此外后者还新增了一个场景校正因子对于大城市中心地区来说，路径损失是直接增加 3dB 的。

➢**其他修正因子**：以上 cost231-hata 模型和 okumura-hata 模型关于因子的校正，主要集中在接收天线的有效高度上，但通过阅读文献，我们发现，两者还有很多其他的修正因子，且这些修正因子在两者模型上的表现是相同的，以下作出具体解释：

•K_{street} ,街道修正因子：设传播方向于街道夹角为θ,

$$K_{street} = \begin{cases} -(-5.9 + 11lgd/6)sin\theta - (7.6 - 10lgd/6)cos\theta & d \geq 1 \\ -(-5.9sin\theta + 7.6cos\theta) & d < 1 \end{cases}$$

街道效应一般在 8-10km 后会消失，所以只考虑 10km 范围之内

•Q_r,准开阔地区校正因子：在 okumura-hata 模型中考虑到了开阔地区的校正因子，而准开阔地区则是在开阔地区校正因子Q_o的基础上再增加 5.5，即：

$$Q_r = Q_o + 5.5$$

•R_u,农村校正因子：

$$R_u = -(lg(f/28))^2 - 2.39(lgf)^2 + 9.17(lgf) - 23.17$$

•K_h,丘陵地校正因子：

$$K_h=\begin{cases}0 & \Delta h<15\\ -(-5.7+0.024\Delta h+6.96lg\Delta h)+7.2-9.5lgh_1 & \Delta h\geq 15,h_1>1\\ -(-5.7+0.024\Delta h+6.96lg\Delta h)+7.2 & \Delta h\geq 15,h_1\leq 1\end{cases}$$

其中Δh表示的是地形起伏高度差，如图所示：

图 2.丘陵地区地形起伏高度示意图

如图所示，从移动台算起，网基站方向延伸 10km,在次范围内地形起伏高度在 10%-90%之间的差值，前提是又多次起伏，我们合理认为起伏次数最少不低于三次。

$$h_1=hmg-\Delta h/8-h_{min}$$

校正因子计算中h_1如上式可进行计算，h_{min}表示的是地形起伏的最小高度

•K_{sp},一般倾斜地形校正因子：

因为斜坡地形可能会有地面的二次反射，如下所示：

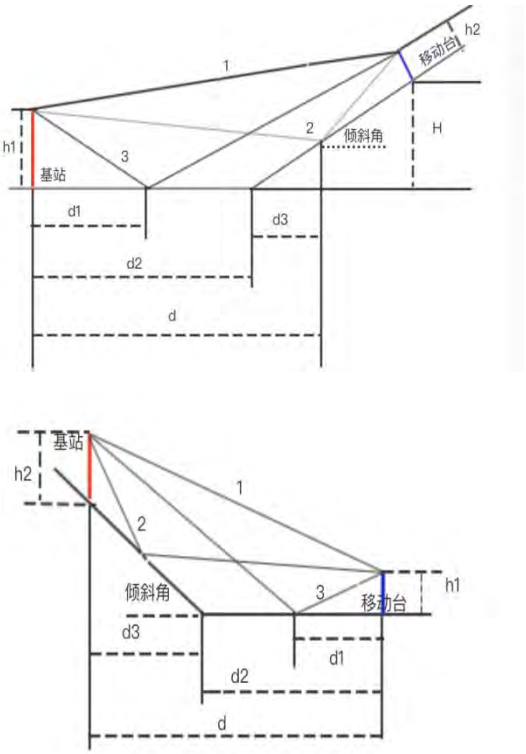

图 3.4.正反斜坡图

近似归纳斜坡地修正因子：

$$K_{sp}=0.008d\theta_m-0.002d\theta_m^2+0.44\theta_m$$

注意到倾斜角在计算时运用的是弧度制，且以毫为单位，而 d 的单位为 km,出于实

际问题考虑，该倾斜角可以近似为移动台前后一公里范围内地形高度的平均倾角。

•K_{im}孤立山峰校正因子：

用刀刃绕射损耗来进行计算，损耗示意图如下：

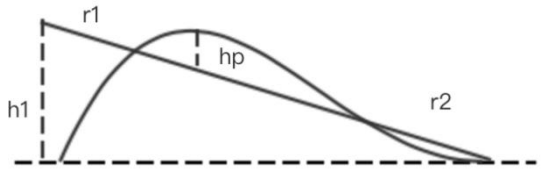

图 5.孤立山峰示意图

需要求解如上图的四个参数以及波长，因此处计算与下述绕射计算较为相同，因此将在后文进行详细阐述以上介绍了部分修正因子的计算的同时展现了修正因子对路径损耗带来的影响，是我们后续设计特征的重要依据。

③SPM 模型：

通过文献[6]，我们了解到在移动网络建设中最常用的模型之一除了上述两者以外还有 SPM 模型，该模型主要适用频率在 800M-2000M 之间，收发距离在 0.2-5km 之间，基站天线高度在 4-50m 之间，移动天线高度在 1-3m 较为适合，而该模型本身涉及到非常多的系数，这些系数在不同城市不同地区都不相同，具体模型表现如下：

$PL = K_1 + K_2 * logd + K_3 * h_r + K_4 * h_t + K_5 * Diff + K_6 * logd * logh_t + K_7 * L_{clutter}$

$K_1 \sim K_7$是在不同场景的不同参数[10]，因此在此我们不将其细化作为特征，其中$Diff$表示的是传播路径中的绕射损耗，$L_{clutter}$表示为地物损耗。

绕射损耗因子与绕射常数有关的具体计算方法如下：

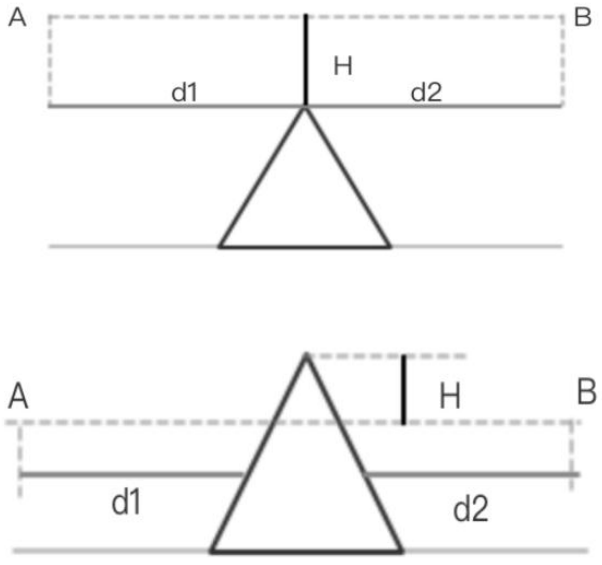

图 6.绕射示意图

如上图两种情形下，第一种是指高 H 处的视距路径无障碍，第二种表示在电波传输过

程中有障碍物。因此，在前者环境下，我们假设障碍物高度为复数，而后者障碍物高度为正数。绕射损耗 Diff 主要与绕射常数有关，常数计算公式如下[6]:

$$v = -H\sqrt{\frac{2}{\lambda}(\frac{1}{d_1}+\frac{1}{d_2})}$$

不同绕射损耗近似值由下式可得[6]:

$$F = \begin{cases} 0 & v \geq 1 \\ 20log(0.5+0.62v) & 0 \leq v < 1 \\ 20log(0.5e^{0.45v}) & -1 \leq v \leq 0 \\ 20log(0.4-\sqrt{0.12-(0.1v+0.38)^2}) & -2.4 \leq v \leq -1 \\ 20log(-0.225/v) & v < -2.4 \end{cases}$$

在利用该模型进行路径损耗计算时，根据收发天线之间的距离以及与障碍物高度的之间的关系，可以分为视距与非视距两种情况，两者的判断条件如下：发射点和接受点之间障碍物的高度没有超过两点一线，则认为是视距，否则认为为非视距。在这两种情况下，算法也是不同的。

1.视距算法

收发天线两点一线意味着中间没有任何遮挡物，因此，绕射带来的损耗以及地物带来的损耗相当于 0，此时的传播模型主要依赖于 自由空间下的传播模型，具体公式如下：

$$PL = K_1 + K_2 logf + K_3 logd$$

k1,k2,k3 可以设置为默认值，通常情况下为 32.44，20，20

2.非视距算法

非视距算法即在介绍模型一开始提到的标准公式。

该标准公式一般运用于半径大于 1km 的小区，所以当路径损耗区域范围小于 1km 时，可能会带来较大的误差，因此作出一定调整。

近点，在$d < d_0$的情况下

$PL_{near} = K_1 + K_{21} * logd + K_3 * logh_r + K_4 * Diff + K_5 * logd * logh_t + K_6 * h_r + K_{clutter} * f(clutter)$

远点，在$d \geq d_0$情况下：

$PL_{far} = PL_{near} + (K_{21} - K_{22}) * logd_0$

其他传播模型，可以详见相关参考文献。[8.9]

5.2 特征设计

5.2.1 设计逻辑

上节中，我们主要对于现有的传播模型及其改进模型进行简单介绍，同时注意到在利用这些模型计算的过程中，涉及到的参数例如有效接收天线高度等因素势必会对最后的损失值造成影响，因此，出于想利用设计的特征达到最后计算损失率的目标要求，希望能得到尽可能全面的特征，在设计特征的初级阶段，我们将上述这些模型在计算过程中出现的参数都称为**涉及特征**，就以 Cost 231-Hata 模型中出现过的参数为例，作为未来我们考虑提取有效特征时的一部分素材,需要注意的是因为数据集本身的数据属性的限制，有些特征是主观意义上进行判断，并非完全保证对最后结果有效，同时也有可能是利用当下数据集无法估算的特征，这一点将在下一节进行主要呈现，主要包括特征判别，提取和构建过程，在本节中做提及的涉及特征仅处于初步设计阶段，不考虑是否能够完成。

在设计初期，搭建整体特征体系时，出于对完整性和多方向性的考虑，我们将特征主要分成五个部分：**涉及特征，衍生特征，已有特征，计算特征，校正特征**。关于涉及特征

的具体定义参考上述表达，而**已有特征**主要参考数据集，数据集主要有三部分构成，工程参数，地图数据，RSRP 标签数据。因 RSRP 是作为标签数据，因此此处已有特征仅包含工程参数和地图数据两部分。**衍生特征**，即通过我们已有数据集，通过构建地图上的三维位置图等，通过几何位置等进行衍生计算而得到的特征数据，**校正特征**则是表示校正因子，即在 cost231-hata 模型下以及奥村模型，okumura-hata 模型，SPM 模型下所涉及到的校正因子。**计算特征**即是我们运用上述 cost231-hata 模型计算 PL,RSRP 的同时计算它与实际给出的 PL,RSRP 值之间的差值,因为不能以一个模型计算来覆盖整体表现，因此我们将 cost231-hata 模型作为主要模型，其余两个最常见模型仅用于计算路径损耗，即计算奥村哈特模型以及 SPM 模型下的路径损耗 PL 值。RSRP 差值在一定程度上可以反映 RSRP 的表现情况,同理 PL 值更是能直观表示路径的损耗情况。计算特征的构建主要是由于上述四个特征都是由理论上进行分析，以理论和模型为依据出发，而我们是可以通过已有部分数据以及学习相关模型粗略进行计算的，以我们可计算得到 RSRP 值为例，与实际 RSRP 标签数据进行比较，不得不承认，是可以通过差值来反映 RSRP 的整体表现，即是一种以结果为出发点和依据，反向思考这差值对于 RSRP 本身的影响，因此考虑通过正反向逻辑完善特征体系的方向性。

同时，在上述特征定义时，需要注意的有四点：

(1) 因为涉及特征本身的计算问题，我们在本节不对其进行计算，但是依据现有模型以及数据集，是可以对衍生特征，计算特征进行计算的，理解这一步是后续做特征的基础。

(2) 在运用 cost231-hata 模型计算时，因为它面对不同传播环境下修正因子不同，我们采用适用于中小型城市的传播模型校正因子。因为，在实际运算中我们发现，当利用大城市或者乡村环境下的校正模型时，达到的 RSRP 与实际 RSRP 做差值时，得到的误差普遍大于我们利用中小型城市模型计算的得到差值，因此我们有理由认为利用中小型城市校正因子来进行计算是在大部分情况下符合要求的，我们也有理由认为该模型在一定意义上是较为可靠的模型。

(3) 在计算过程中我们可以得到相对高度，但相对高度有部分数据表现为小于 0，因此我们可以找到相对高度小于 0 所对应的传播环境，该相对高度小于 0 的地物分布的大致类型可以说明地物类型对我们目标的影响，而地物类型会影响我们对于传统模型的选择，因此可以挑选适用于这一模型中所涉及的参数作为我们的特征，很明显这一类特征属于涉及特征，为方便起见，我们就直接将当代主要研究的三种模型中涉及到的所有参数都作为我们设计特征的一部分。

(4) 校正特征是的是将所有校正因子都归为一类，需要与之前的涉及特征（模型中涉及到的参数）加以区分。

因此，为了更清晰的表达我们指标设计的观点，将其定义用表格化的方式呈现。

表 4.特征定义

特征类型	定义
已有特征	数据集中除 RSRP 标签数据外的所有数据属性
涉及特征	已有传统模型中涉及到的相关参数
衍生特征	通过已有数据能通过几何等数学手段得到的其他衍生数据
计算特征	通过传统模型相关计算得到 PL,RSRP 值及其与实际 PL,RSRP 值之间的差值
校正特征	cost231-hata 模型，奥村模型，okumura-hata 模型，SPM 模型中所有校正因子

5.2.2 设计特征合理性验证

根据上述指标，关于其中衍生指标的计算示意图如下：

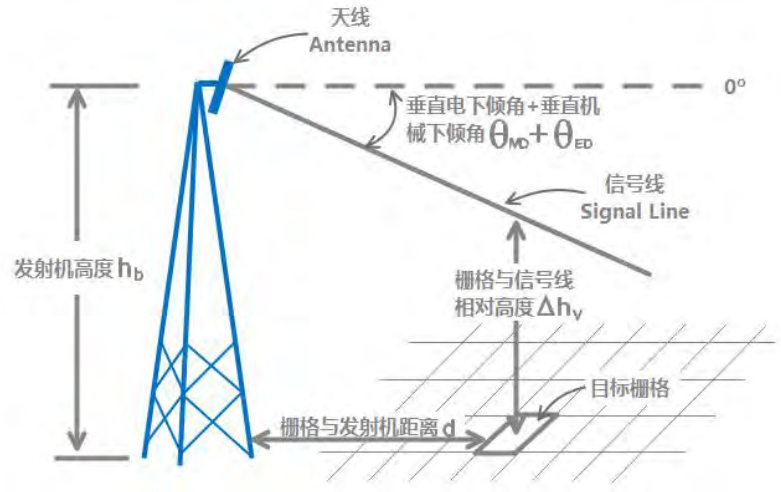

图 6.相对高度示意图

如图所示，通过发射机高度，栅格于发射机距离，垂直电下倾角和垂直机械下倾角可以计算得到栅格与信号线的相对高度。该相对高度即作为我们通过已有数据得到的衍生数据，将其作为一种特征，即是属于我们衍生特征的一部分，通过小区 234101 的相对高度计算，我们找到其中相对高度小于 0 的地物类型分配如下：

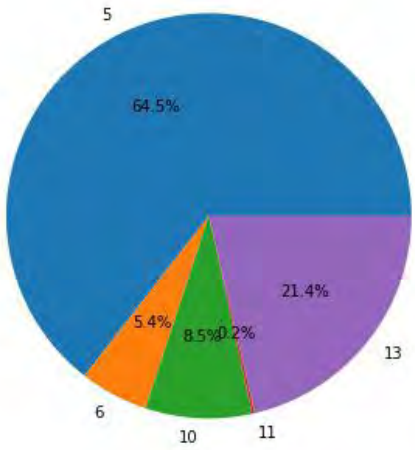

图 7.相对高度小于 0 的对应地物类型分配图

通过上图，我们可以明显观察到相对高度这一特征小于 0 时，主要的地物类型是市区开阔地带，城区小高度大密度建筑群，具体地物类型如下表所示：

表 5.相对高度小于 0 的对应地物类型

数字标号	5	13	10	6
地物类型	市区开阔区域	城区<20m 高密度建筑群	城区 (>60m) 超高层建筑	道路 开阔区域
百分比	64.5%	21.4%	8.5%	5.4%

市区开阔区域和城区内小于 20m 的高密度建筑群构成了其中 85.9%的分配比例，可以

认为这两者占据了主要部分，因此，有理由认为关于相对高度这一特征与地物类型之间存在一定相关性，因此可以作为影响计算 RSRP 的指标之一。

关于计算特征，我们以 cost231-hata 与相对高度结合的模型为例，得到 234101 的下列部分 RSRP 为例:

表 6 小区 234101 的部分 RSRP 值与实际 RSRP 对比

计算 RSRP	-101.01	-90.74	-97.64	-91.11	-90.94	-93.91	-94.05	-101.13
实际 RSRP	-110.70	-117.20	-112.20	-105.70	-115.63	-116.20	-124.70	-108.70
差值	-9.69	-26.46	-14.89	-25.09	-30.76	-15.79	-18.15	-4.57

从上述数据中，我们发现，该模型与实际 RSRP 之间的存在一定的差距。而这一差距也能反应该模型对 RSRP 的反应表现，即综合表现了模型中涉及到的参数对 RSRP 的影响，势必会成为我们研究特征时重要的一部分。

5.2.3 整体特征体系

因此，通过上述粗略的计算及分析，我们有理由相信可以通过以下这些特征来进行最后的模型建立，设计的具体指标如下：

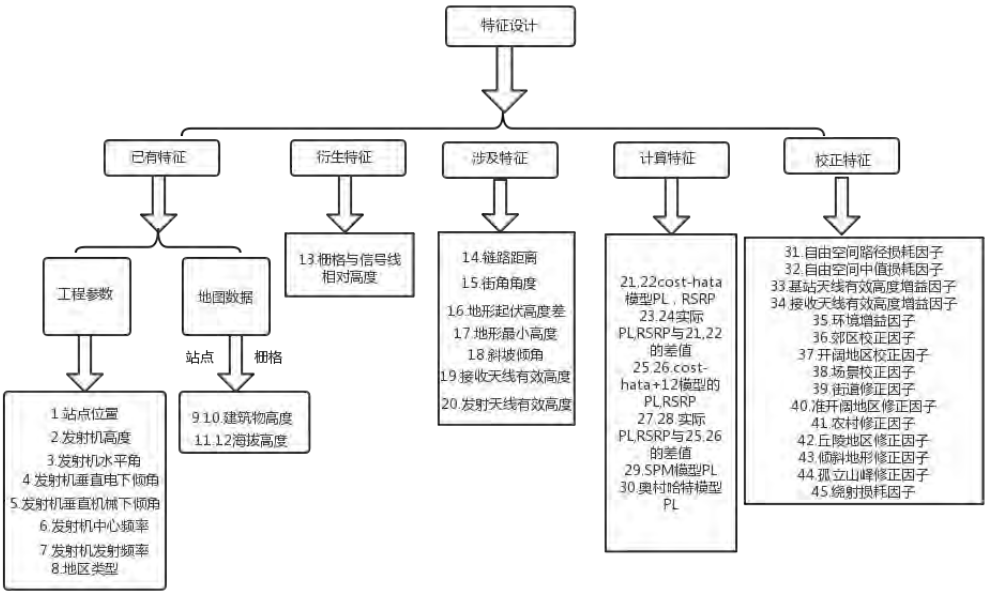

图 8.整体特征体系设计图

通过阅读大量相关文献，我们发现在不同文献中，不同模型下对于同一参数的表达都各有不同，因此，我们在此出于规范化表达的考虑，统一将这些参数以上图形式进行规范化表达，例如涉及指标中的链路距离，即是指栅格与发射机之间的距离，移动天线高度或是接收天线高度均表示为用户天线高度，场景修正因子即使在固定模式下是常数，我们仍然将之作为校正特征中的一部分用以体现不同环境对于传统模型的影响。工程参数下，机械下倾角是通过调整天线面板后面的支架来实现的，是一种物理信号下倾；而电下倾角是通过调整天线内部的线圈来实现的，是一种电信号下倾。实际的信号线下倾角是机械下倾角和电下倾角之和。以上，通过对于不同模型的解释以及对各个模型中的参数和修正因子包括计算结果与实际 RSRP 值之间的差值等均做了详细的定义搭建了整体特征体系，为后续特征提取和最终构建，排序等工作奠定基础。

六．问题 2:基于设计特征的特征筛选及量化排序

基于完成了上述整体特征体系的构建，考虑到特征体系内由于我们主观判断或是特征冗余等等会给最后机器学习技术带来效率低下的影响，因此我们需要对特征进行筛选，提取，等工作。首先通过已有数据判断是否可以进行特征的计算，再对通过计算后得到特征进行数据预处理，包括数据过滤，清洗等步骤，然后对得到的特征数据先进行发散性分析，具有一定发散程度的特征数据进行相关性分析，不光是特征与特征之间的相关性分析，还包括了特征与目标之间的相关性分析，通过相关性进行量化排序后最终确定需要从体系中提取出的特征，构建真正意义上机器学习中需要运用到的特征平台。

6.1 特征筛选

在数据筛选阶段，首先我们考虑在已有特征体系下进行特征过滤，即通过已有数据无法进行计算得到的特征进行剔除，剩下可以通过一系列计算得到的可达特征。再对这部分可达特征进行数据计算得到可达特征数据，数据计算即进行数据预处理，包括数据过滤，填充等等，针对这些特征数据进行离散数据分析，选择有一定离散表现的数据作为特征数据，再进行相关性分析，最终选择相关性较高的数据作为我们需要运用到最终模型中去的相关特征。整体特征选择的思路如下：

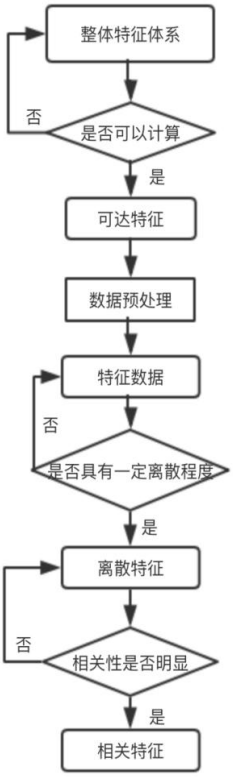

图 9.特征筛选的整体流程

以上为我们特征提取工作的整体思路。其中数据预处理模块，出于对于实际数据集考虑，主要进行数据过滤，数据填补等工作。

针对于数据集中已包含的数据，我们将数据过滤规则表示如下：

1.剔除采样数据距离基站过远或过近的样本数据

2.剔除信号覆盖率过于低的样本数据

样本数据距离基站距离即可用收发距离 d 进行判别，过低信号则可用 RSRP 标签数据进行判别。

针对于数据异常值填补工作，我们通过阅读相关文献，将规则如下所示：

1.对于基站高度小于 0 的基站来说，将其赋值为 50

2.在相对高度小于 0 的情形下，将其相对高度赋值为 1

在完成上述数据准备工作下，需要明白当我们面对不同环境下进行计算时，部分因子随之改变，因此我们在计算 PL,RSRP 值时是需要进行分类讨论的：

1.对于栅格位置与基站位置高度一样的数据，我们认为此类数据表明该水平面无建筑物，可以认为收发天线之间的距离为 0，对于此类数据进行两种模型的计算，利用 SPM 模型和 cost231-hata 模型中对应校正因子的分别进行 PL 值的计算。

2.当遇到不同环境下电波传播时，需要在对应环境下计算校正因子,绕射损耗值

因此经过上述数据处理后，考虑到 SPM 模型与 cost231-hata 在实际应用中的采用情况，运用范围等因素，我们得到了以下 22 个特征，并将之归纳为两个部分：

表 7.特征数据

已有特征	计算特征
1.小区所属站点的栅格位置，X 坐标	16.Cost231-hata 计算得到的 PL
2.小区所属站点的栅格位置，Y 坐标	17.Cost231-hata 计算得到的 RSRP
3.小区发射机相对地面的高度	18.绕射损耗
4.小区发射机水平方向角	19.不同环境下校正因子
5.小区发射机垂直电下倾角	20.自由空间传播损耗
6.小区发射机垂直机械下倾角	21.基站与目标栅点直线距离
7.小区发射机中心频率	22.SPM 下路径损耗
8.小区发射机发射功率	23.理论上传播路径损耗值
9.栅格(X,Y)上的建筑物高度	24.实际 RSRP 值
10.小区所在栅格的建筑高度	25.相对高度
11.栅格(X,Y)上的海拔高度	
12.小区所在栅格的海拔高度	
13.栅格位置 Y	
14.栅格位置 X	
15.地物类型	

上表中值得注意的主要有三点：

1.上表中所展示的特征与问题一中列出的整体特征体系是不同的。可以说问题一中的体系包含了上表中所有特征，而上表中所有特征也能反应整个体系的框架设计逻辑。因为 SPM 有着更广泛的适用性，因此表中 20 与 15 可以用于不同中情景下进行计算，同时表中 17 虽称为较正因子项，但内在又包含了郊区校正因子，开阔道路校正因子，准开阔道路校正因子，农村校正因子，城市校正因子这 5 种校正因子，这些因子如之前数据准备中所说是需要在分类讨论的前提下进行运用的，在某种特定场景下进行预测时，只可能用到其中一种，因此在此我们也是用一种特征来进行泛化表征。

2.上述特征值是可以进行具体计算的，为了方便理解，我们把问题一中的所有特征比作一个“基因库”，而此处特征称作“携带基因”，并且此处特征已经不属于 cost231-hata 模型下的“衍生品”，而是出于对更好 AI 传输模型的要求，挑选的最泛化且简便，综合性高的特征，例如其中 SPM 下路径损失，可记作 SPM-PL。

3.关于地物类型[11]，数据本身只是给出了不同地物类型的索引标签，但实际上地物类型是影响传播模型的关键，是一个不可以剔除的指标，我们需要将其量化。根据阅读相关文献，我们发现对于这一类数据来说，通过重要性进行排序后给予一定权重是解决这个问题的关键，即可将将索引转化为数据表示。

完成了上述对于特征数据的提取，接下来需要完成的是对于这些数据特征进行筛选和

提取，构建我们后续模型需要用到的特征平台。

6.2 特征构建

搭建真正的特征平台时，我们需要对上述 25 个特征进行发散性分析以及相关性讨论，从而挑选出真正适用于模型且有利于提高模型真正效率的特征，并将这些特征作为后续机器学习的主要依据。为了对其相关性进行研究，我们可以通过量化排序的方式，直观且清晰的表达对最终特征的选取，根据图九所示流程图，将过程主要分成两个部分，一则先进性发散性分析，二则进行相关性分析，具体操作如下：

（1）发散性分析：

对于上述 25 个特征而言，因为对于地物类型是通过重要性进行权重排序的，所以在此不需要对其进行发散性分析。除此之外，因为相对高度数据即使离散程度很小也被我们认作是重要特征，因此在此也不进行离散性分析。基于上述考虑。首先，我们对其剩余特征数据进行整体性数据观察，关于数据本身的极差，中值，方差等并将其归一化后根据方差大小进行排序，以 2304101 小区为例得到相关数据如下：

表 8.特征数据离散程度

排序	特征	均值	极差	中值	方差
1	栅格海拔高度	0.9018	0.9013	0.0004	0.0203
2	Cost231-hata-PL	0.9132	0.9112	0.0006	0.0244
3	栅格建筑物高度	0.0118	0.0000	0.0018	0.0418
4	自由空间损耗	0.8041	0.7913	0.0068	0.0827
5	理论传播损耗	0.4784	0.4794	0.0121	0.1100
6	RSRP	0.5067	0.5053	0.0130	0.1141
7	发射机中心频率	0.01745	0.0000	0.0135	0.1162
8	SPM -PL	0.4932	0.4912	0.0170	0.1305
9	栅格位置 Y	0.3822	0.3845	0.0188	0.1372
10	基站栅格位置 Y	0.3689	0.3712	0.0200	0.1415
11	基站海拔高度	0.3917	0.3846	0.2038	0.1428
12	栅格位置 X	0.5494	0.5656	0.0239	0.1547
13	发射机相对地面高度	0.3650	0.3692	0.0240	0.1549
14	基站所在栅格建筑物高度	0.0760	0.0000	0.0246	0.1570
15	基站栅格位置 X	0.5395	0.5565	0.0283	0.1682
16	小区发射机垂直机械下倾角	0.3845	0.3571	0.0298	0.1725
17	小区发射机发射功率	0.4953	0.3571	0.0320	0.1789
18	小区发射机垂直电下倾角	0.4249	0.5000	0.0381	0.1953
19	基站与目标栅点直线距离	0.1356	0.0456	0.0543	0.2330
20	绕射损耗	0.8199	1.0000	0.0546	0.2337
21	小区发射机水平方向角	0.4957	0.5000	0.0842	0.2902
22	校正因子	0.2318	0.0025	0.1656	0.4070

从上述数据中，因为每个数据属性的量纲不同，因此用归一化的手段进行方差分析观察数据的离散程度，我们发现上述所有特征的方差表现最为剧烈的是校正因子，而相反的栅格所在的海拔高度以及运用 cost231-hata 模型计算出的路径损耗值的方差较小，最大与最小方差之间存在着 40 倍的差距，因此我们考虑在后续分析中剔除掉方差过小的特征，

因为此类特征在不同情况下差异性不明显，侧面反映这个特征对传播模型的影响并不大。

(2) 相关性分析：

相关系数，即是考察变量与变量之间的相关程度，当相关系数的绝对值越大，说明两者间的相关性越强，通常情况下，我们将相关系数划分为以下五个部分：

表 9.相关强度表

相关强度	极强相关	强相关	中等相关	弱相关	极弱（无）相关
相关系数范围	0.8-1	0.6-0.8	0.4-0.6	0.2-0.4	0.0-0.2

现代统计学中最常见的三大相关系数分别为：皮尔逊系数，斯皮尔曼相关系数，肯德尔系数，但由于数据本身的不同，系数求法的适用范围不同，这三种系数也有不适应计算的情况发生，因此，本文在此用三种不同的相关系数求法分别对上述 22 个特征进行相关系数的计算，具体如下所示：

1.皮尔逊相关系数：

皮尔逊系数适用于两个变量之间是线性关系，且是连续数据分布，变量总体分布是正态分布或者接近正态分布的单峰分布，同时每个变量的观测值是成对出现的，每一对数据应该是相互独立的。

2.斯皮尔曼等级相关系数：

斯皮尔曼系数对数据要求并无皮尔逊那么严格，他只需要两个变量观测值成对出现即可，也可以用是通过其他连续数据资料转化而来的等级资料，换句话说斯皮尔曼系数可以看作是两个皮尔逊系数经过排行后得到的等级系数。

3.肯德尔等级相关系数：

肯德尔等级系数是用来测量两个变量的统计依赖性，取值范围在-1 到 1 之间，当取 1 时，说明两个变量之间拥有一支的等级相关性，当-1 时，说明两个变量之间拥有完全相反的等级相关性，若取值为 0，说明两者相互独立。而该相关系数对数据的条件要求与斯皮尔曼相关系数保持一致。

基于上述不同相关系数对于不同变量关系之间的考察，我们得到如下相关系数图式：

➢皮尔逊相关系数：

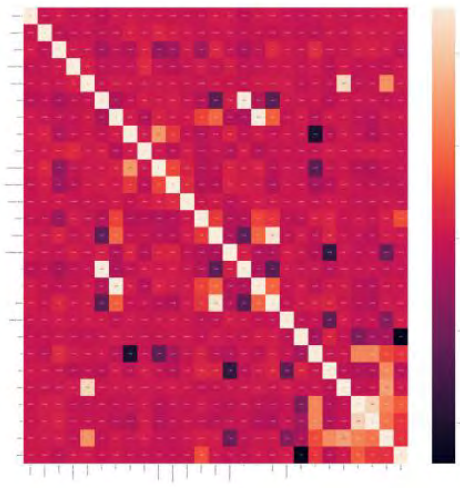

图 10.皮尔逊相关系数热力图

将上述相关系数与目标之间的相关性进行排序，如下所示：

表 10.皮尔逊相关系数排序

排序	特征名称	相关性	排序	特征名称	相关性
1	RSRP	1	12	栅格上的海拔高度	0.0346
2	理论 PL	0.9748	13	栅格的海拔高度	0.0257
3	自由空间传播损耗因子	0.3290	14	校正因子	0.0254
4	收发距离	0.1800	15	小区发射机中心频率	0.0089
5	Cost231-hata-PL	0.1545	16	站点的栅格位置，X	0.0064
6	SPM-PL	0.1494	17	发射机水平方向角	0.0055
7	绕射损耗	0.0891	18	栅格位置，X	0.0049
8	建筑物高度	0.0418	19	发射机垂直机械下倾角	0.0031
9	小区发射机发射功率	0.0398	20	发射机垂直电下倾角	0.0031
10	栅格位置，Y	0.0360	21	发射机相对地面的高度	0.0020
11	站点的栅格位置，Y	0.0353	22	站点栅格建筑物高度	0.0011

通过上表我们发现，关于已有特征，他们普遍与目标相关性较小，而我们给出的计算特征，因为是通过已有数据进行一系列综合计算得到，总体上而言表现出来的相关性要较原有数据大的多，例如自由空下的损耗因子，例如运用特定模型下的 PL 值，收发天线距离等。

说明综合性特征的表现效果要比单独特征表现效果好的多。皮尔逊系数本身计算即是一种“线性”关系的体现，因此，我们不能单一的只用皮尔逊系数进行相关性判断，而是需要借助其他层面的相关性进行进一步判别。

➢斯皮尔曼相关系数：

同理，我们可以得到关于此相关系数的热力图：

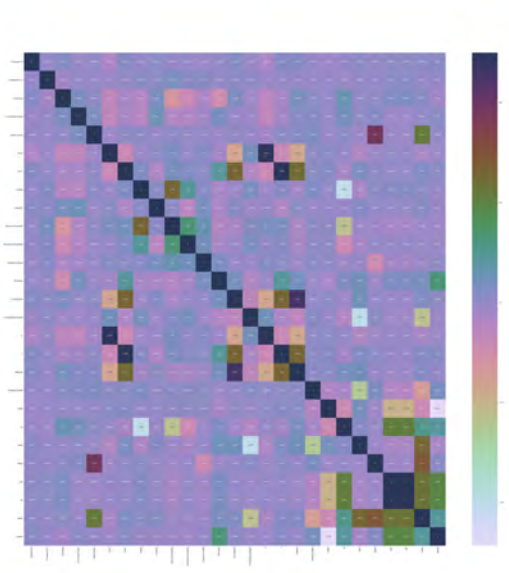

图 11.斯皮尔曼系数热力图

并给出相应排序，如下：

表 11.斯皮尔曼相关性排序

排序	特征名称	相关性	排序	特征名称	相关性
1	RSRP	1	12	栅格位置，Y	0.0274
2	理论 PL	0.9733	13	栅格的海拔高度	0.0191
3	自由空间传播损耗因子	0.3873	14	栅格位置，X	0.0161
4	收发距离	0.3873	15	站点的栅格位置，X	0.0151
5	Cost231-hata-PL	0.1653	16	站点所在栅格海拔高度	0.0112
6	SPM-PL	0.1597	17	发射机垂直机械下倾角	0.0097
7	绕射损耗	0.0776	18	发射机垂直电下倾角	0.0080
8	建筑物高度	0.0737	19	站点栅格建筑物高度	0.0072
9	小区发射机发射功率	0.0459	20	发射机中心频率	0.0066
10	校正因子	0.0291	21	发射机水平方向角	0.0012
11	站点的栅格位置，Y	0.0278	22	发射机相对地面高度	0.0004

因为斯皮尔曼本身求解系数的方法是基于“秩”检验，因此，在一定程度上排除了原数据过大过小而带来的影响，比皮尔逊运用范围更广，是一种较好的“重要性”等级关系的体现。与皮尔逊系数计算相比，在上述特征中，前 9 项的排名顺序不变，有理由认为这部分特征属于我们的重要特征，且相关性更为明显，在相关性低于 0.03 的情况下，后续特征值原本也不会对目标值产生过大的影响，说明该系数检验前后保持了一致性，且该方法在总体上而言相关性比皮尔逊略高，也说明了数据本身不足够满足皮尔逊条件。

➢**肯德尔相关系数:**

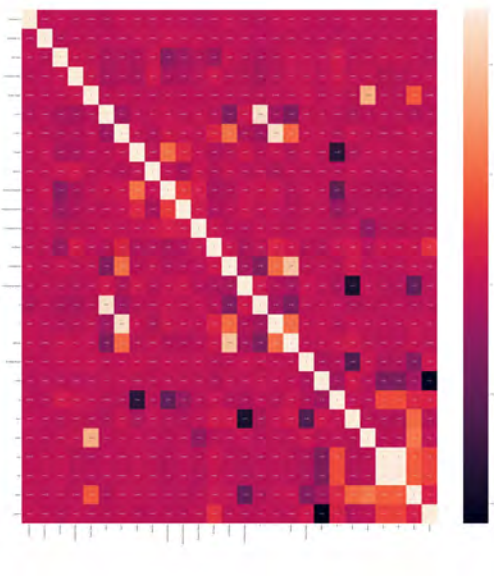

图 12.肯德尔系数热力图

因为基于上述检验过程中出现的问题，为了得到对于我们特征数据更为合适的相关性表现，我们考虑用肯德尔系数进行计算，它除了降低了皮尔逊系数对数据要求的条件外，即适用于斯皮尔曼系数计算的适用范围，一定程度上也是等级相关性的体现。同理得到排序如下：

表 12.肯德尔相关性排序

排序	特征名称	相关性	排序	特征名称	相关性
1	RSRP	1	12	栅格位置，Y	0.0184
2	理论 PL	0.8718	13	栅格的海拔高度	0.0129
3	自由空间传播损耗因子	0.2612	14	栅格位置，X	0.0109
4	收发距离	0.2612	15	站点的栅格位置，X	0.0102
5	Cost231-hata-PL	0.1097	16	站点所在栅格海拔高度	0.0075
6	SPM-PL	0.1068	17	发射机垂直机械下倾角	0.0068
7	建筑物高度	0.0581	18	发射机垂直电下倾角	0.0060
8	绕射损耗	0.0576	19	站点栅格建筑物高度	0.0055
9	小区发射机发射功率	0.0340	20	发射机中心频率	0.0053
10	校正因子	0.0229	21	发射机水平方向角	0.0007
11	站点的栅格位置，Y	0.0187	22	发射机相对地面高度	0.0004

我们发现，肯德尔系数排序后与斯皮尔曼系数排序的结果几乎一样，但相关性程度比斯皮尔曼排序有一定程度的减弱，但前九项仍然是一个相关性的分界点，但是此相关性较三种系数而言是最小的，此问题发生的原因在于该系数通常情况下主要用于检验判断一致性，可以说是一种“方向性”等级排序，因为传播路径，传播环境的复杂性，在方向上变化不能保持较高相似性是可以理解的。

同时需要注意的是，我们在一开始初始特征设立有 25 个，但进行离散性分析和相关性分析的只有 22 个，其余三个特征分别为地物类型，cost231-hata-RSRP 以及相对高度，第一个特征需要我们在模型建立时设置权重，对其无论进行相关性分析或是发散性观察都没有意义。后两者则是我们作为最后模型训练中的基础重要特征，因为根据问题一我们可以发现相对高度的计算出现在各大校正因子内，且 cost231-hata-RSRP 值是作为一个重要参考指标进行对照，因此在离散化下无需对其进行观察。只需单独对他们进行相关性检验，如下：

表 13.单独相关性分析检验特征

特征	相对高度	Cost231-hata-RSRP
相关性	0.01657	0.0112

与之前的集中特征相比，这两个特征值与目标之间的相关性总体而言并不弱，因此可以作为后续模型训练中的部分特征。

七．问题 3:建立关于 AI 传播模型的 RSRP 预测模型

依据本题的要求，需要以 Tensorflow 框架作为我们模型运行的基础，基于 Tensorflow 的定位主要是深度学习库，且我们本身数据量非常庞大，并且需要我们最终的计算精准度较高，所以采用深度学习的方法对数据本身进行提炼，选取合适特征.首先我们在上一问题中涉及到的 25 个特征中，因为地物类型这一特征的特殊性，考虑通过 onehot 独热编码进行计算，但是通过阅读相关文献，我们发现处理这一类型的数据时，较多的都依赖于不同地物类型的重要性排序，因此，针对不同地物类型，我们根据相关因子也对其进行排序，再放入模型。同时需要注意的是，因为栅格位置以及站点所在栅格位置数据本身的局限性，所以在套入模型中时，我们人为的将其剔除。最终我们选择了 15 个特征放入模型。

7.1 神经网络模型

7.1.1 网络结构比较

作为目前非常火热的研究方向之一——多层神经网络作为我们考虑的模型之一。首先我们对神经网络进行介绍。首先，我们需要以一个经典神经网络模型作为参考，一个包含三个层次的神经网络结构如下：

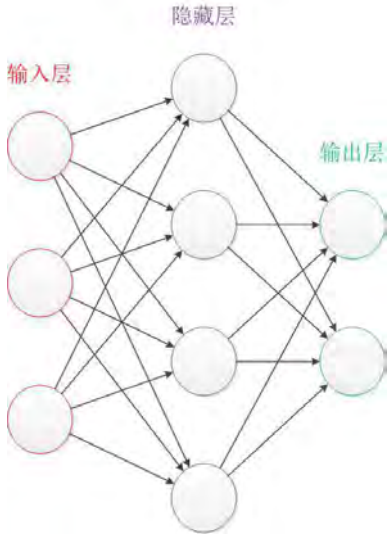

图 13.神经网络结构示意图

如图所示，该神经网络结构呈现方式为（3，4，2），即表明 3 个输入层，4 个隐藏层，2 个输出层。

需要注意的主要有以下三点：

1.在设计神经网络模型时，往往输入层和输出层的个数是固定的，中间部分的隐藏层是需要我们进行设计的。

2.图中箭头表示我们的数据流传输方向

3.图中圆圈表示的是神经元，而不同圆圈间的连接线对应的是不同权重，而权重的确定是通过模型训练得到的。

以下对于上述神经元做简单介绍：

神经网络本质意义上来说，就是模拟了人类大脑对于事物进行判别的过程，，通过生物学研究发现，人类一个神经元通过多个树突来进行信息的传入传出，而一个神经元只有一个轴突，轴突末端有很多轴突末梢可以给其余多个神经元继续进行信息的传递，从而将神经元连接起来，达到信息传递的目的，而该轴突末梢与下一神经元之间进行连接的未知在生物学上被称作“突触”，具体示意图如下：

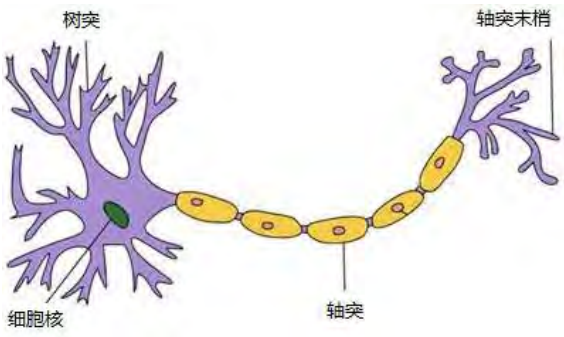

图 14.神经元示意图

若以单个神经元传递为例，即对应的是我们神经网络中的单层神经网络：

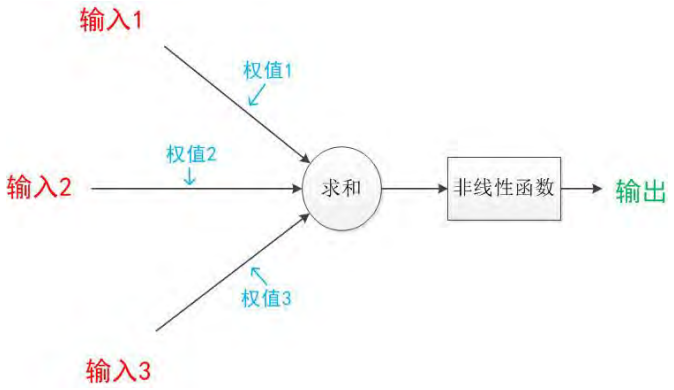

图 15.单层神经网络结构示意图

需要注意的如何设置权重，因为从上述示意图中，我们明显可以发现，输入与输出的关系如下：

$$输出= \Sigma输入\ i*权重\ i(i = 1,2,3)$$

因此，我们将之拓展为深度学习中的多层神经网络，相关结构示意图如下：

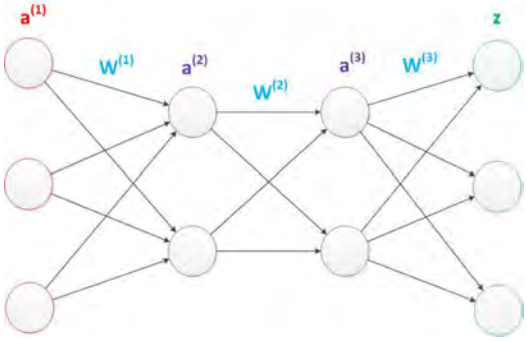

图 16.多层神经网路示意图

与之前单层神经网络进行对比，我们发现，他们之间的差异主要体现在中间隐藏层的增加，而随着中间隐藏层的增加，原有的输出层变为了下一步的输入层，相应的权重个数也在不断往上增加，在图 16 中，可以发现其中需要设置的有 16 个权重，而同样，保证输入层和输出层都为 3 个的基础上，若调整中间层的借点个数，整个网路的权重个数也会随之增加：

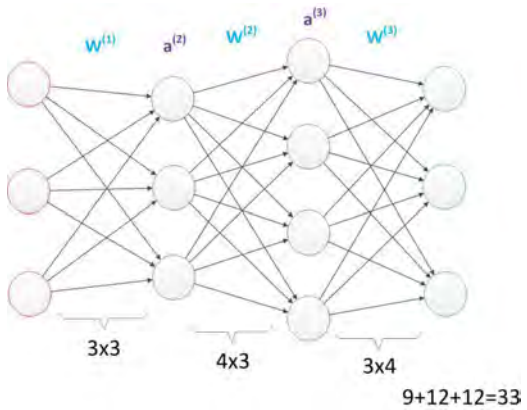

图 17.改变中间节点数的神经网络结构示意图

图 16 与图 17 保证层数不变改变中间节点，有效证明了权重个数与节点数有关。

以下为了更直观表现，我们考虑保持参数不变，改变层数，如下：

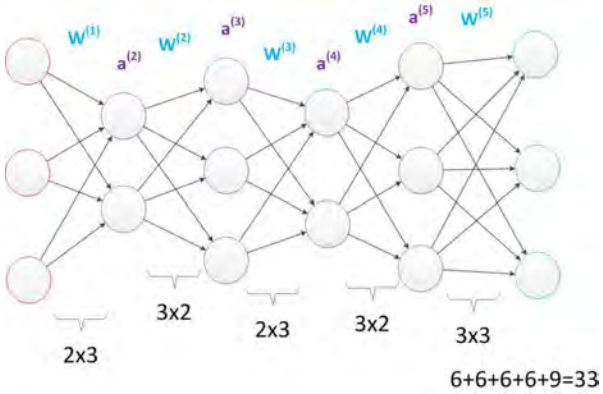

图 18.权重数不变改变层数神经网络结构示意图

图 17 与图 18，同样的 33 个权重参数，说明了，在参数不变基础上是可以保证有更深层次的表达效果的。

7.1.2 激活函数

以上述结构图作为基础，还需要了解的是神经网络常用的激活函数。激活函数就是对各个路径下输入求和之后进一步增强的函数，在上述结构中我们能很明显发现，无论神经网络中间有多少层，输入和输出的都是线性组合，是最原始的神经网路，但是为了更逼近最后的结果，我们需要决定引入非线性的函数作为激活函数，提高神经网络的表达能力。

典型的有如下三个激活函数：

1.Sigmoid 函数：

$$f(z) = \frac{1}{1 + e^{-z}}$$

函数图像与导数图像如下所示：

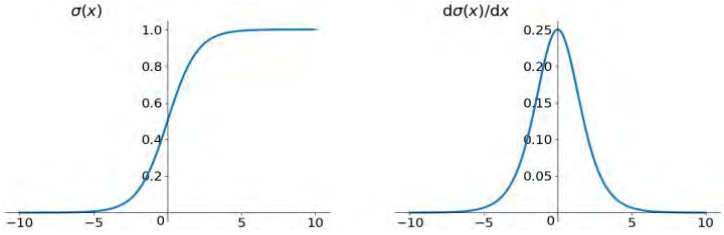

图 19.Sigmoid 函数导数图

2.Tanh 函数：

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

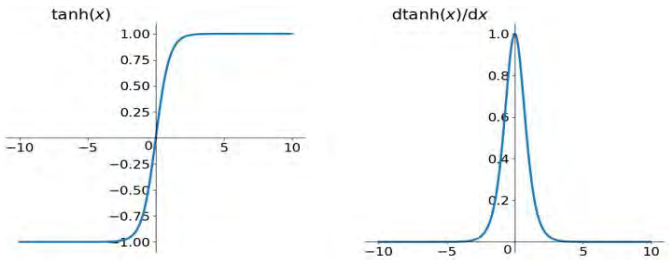

图 20.Tanh 函数导数图

③relu 函数：

$$Relu = max(0, x)$$

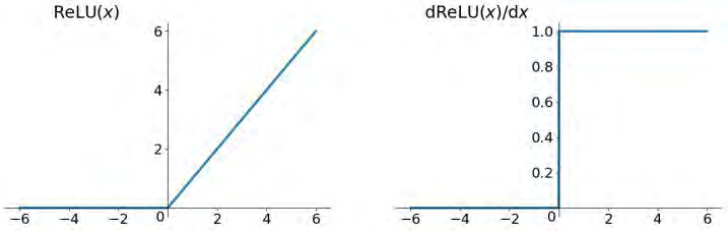

图 21.Relu 函数导数图

通过上述三种函数直观上的对比分析，我们发现，sigmoid 函数是一种便于求导的平滑函数，但是缺点也很明显，在深度神经网络下，运用该函数会让梯度消失发生的概率提升。为了便于理解，譬如说，当我们给初始化的神经网洛的权重是一个（0，1）之间的随机数，若按反向传播的算法，每向前传播时，梯度值会见效为原来的 0.25 倍，那么很有可能当隐藏层较多时，最后梯度值非常接近于 0。若我们给初始值处于一个大于 1 的范围的化，那么就会出现梯度爆炸的情况。同时，在运用该函数时，计算输出均值并不是 0 均值。以及该函数运用的是幂函数，过于耗时，会加大我们模型的训练时间。而 Tanh 函数虽然解决了不是 0 均值的问题，其余两个问题仍然存在。至于 ReLU 函数，在正区间下，他解决了梯度消失问题同时收敛速度和计算速度较前者有大幅提高，但是 0 均值问题仍然需要注意，同时也需要注意有些神经元因为某些可能永远不会被激活。但是总体来说，ReLU 函数是当下最为常用的激活函数。

7.1.3 损失函数及其问题

损失函数是指我们用模型进行预测的预测值与已知实际值之间的差距。在训练模型时，通过损失函数的不断减小从而达到训练处更高准确率的模型效果，常见损失函数有均方差，交叉熵等等。

过拟合：

上述损失函数适用于优化模型训练的一部分，是希望通过它来判断模型的表现好坏的重要参考指标，但是一个模型在过于复杂以后可能会发生过拟合的问题，也就是说，在训练过程中，模型更好的去对噪音数据等进行记忆，而非学习数据集的整体表现，虽然可能在训练数据下出现损失很小甚至为 0 的情况，但是运用到未知情况下，表现就会呈现巨大变化，因此需要采用防止过拟合的手段，通常手段有 early stopping(计算验证集的精准度，当精准度不再提高时停止训练)，正则化，Dropout(修改神经网络本身，通过使部分节点失活提高泛化性)等等。

优化器：

机器学习的本质在于寻找最优模型，即需要我们将损失最小化，通常方法有梯度优化，其中又可以细化为批量梯度优化，随机梯度优化，小批量梯度优化等。但梯度优化本身不能很好的保障收敛性，且在随机梯度优化时，他对所有参数应用相同的学习速率，对于稀疏数据来说，本身我们希望其可以进行大一点更新，所以提出了几个新的优化器，例如 Adam，Adagrad 等，通过阅读相关文献，我们发现实践证明了 Adam 比其他适应性学习方法的学习效果更好。

7.1.4 全连接神经网路及其相关参数

基于上述对于神经网路模型的描述，对于本文研究目标来说，我们采用全连接神经网络模型。全连接神经网络，即按照字面理解，第 N-1 层的神经元与第 N 层的所有神经元之间进行连接，即全连接的含义。

本文主要研究的研究目的是通过该模型进行 RSRP 预测，希望能够在新的环境中快速测定特定地理位置的 RSRP 值，从而可以通过预测得到 RSRP 值来确定无线信号的覆盖强度。根据强度来制定基站位置建设，从而减少建设成本，提高网络传输效率。因此，如何更为精准的预测特定 RSRP 值成为解决问题的核心。我们主要通过 RMSP（均方根误差）来进行判断。因此将之作为输出层。同时结合问题二中得到的特征，排除掉原有的位置参数带来的影响后得到的 15 个特征作为输入层，设置了三层隐藏层，三层隐藏层的节点数分别取做(32，16，8)。

同时选择了 Adam 优化器，Dropout 防止过拟合，Relu 函数作为激活函数，以及均方根作为损失函数。以上的选择都是经过一定综合考虑下得到的具有最优表现的选择。

最终，我们线下划分 200 万数据作为训练集，100 万数据作为测试集，在 tensorflow 框架下，为了将模型更好的进行呈现，我们可以得到关于全连接神经网络结构图：

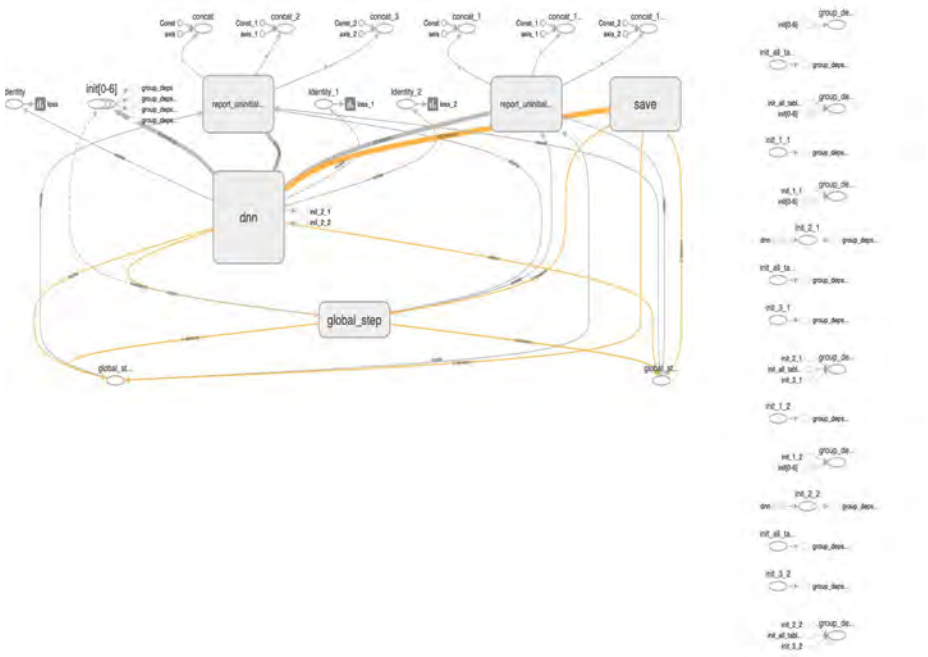

图 22.tensorflow 框架下全连接神经网络结构示意图

通过训练，我们得到该全连接神经网络线下 RMSP 为 9.698。

7.1.5 深度神经网络线性组合回归器（DNNLinearCombinedClassifier）

上述仅是用单个神经网络模型——全连接神经网络（深度神经网洛，DNN）进行计算，出于想要在多个模型下取得最优化效果模型，我们拟采用混合模型进行再次计算，考虑将 DNN 神经网络于线性回归结合起来，在同样训练集和测试集数目下得到线下 RMSP 值为 9.953.该值与前者相比的 RMSP 值反而呈现上升状况，说明与线性回归模型的联合并不会给模型带来增益效果，反而使得 RMSP 效果不如前者，可以说具有一定减弱效果。因此我们对该模型不继续进行线上得分计算。

7.2 随机森林模型及其参数

随机森林是一种有监督的学习算法，就字面意义上而言，可以把该模型拆分为两种类型“随机”“森林”。所谓的“森林”就是指多个决策树的集成。以下对决策树进行简单介绍：

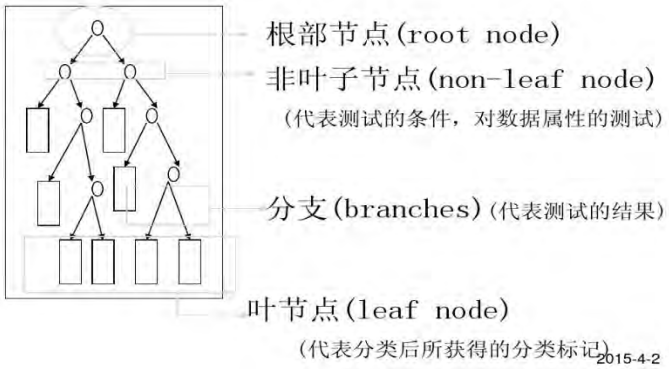

图 23.决策树概念图

随机森林即是通过将多个决策树合并在一起获得更为准确和稳定的预测。随机森林不仅可以用来做分类，也可以用来做回归分析。随机森林中树的增加会给模型带来额外的随机性。决策树中，每个节点被分割成最小化误差的最佳特征，而在随机森林中，我们选择的随机特征来构建最佳的特征分割。所以在实际运用中，我们可以考虑在每个特征下进行随机阈值使得“随机性”更为明显，一定程度上增大了模型的泛化能力。同时，随机森林算法的另一个优势在于它能够测量每个特征对于预测的重要性。对于有部分人提出的随机森林过拟合的问题，可以通过创建随机的特征自己，并通过他们来构建较小的决策树，这样可以防止大部分情况的过拟合，虽然会使得计算速度变慢，但是计算速度还是需要取决于随机森林模型中树的总体数量。

关于我们在运行随机森林时选取的参数如下：

想要建立的子树的数量：20

较多子树会让计算速度变慢，但是会提升模型的精准度，此处参数设计是我们根据日常实际操作得出的较为合适的子树数量。

决策树最大深度：25

在平时，数据量较为小的情况下，这个最大深度值是可以不输入的，但是考虑此处我们的数据量巨大，因此需要用这个最大深度来进行限制。

最小样本叶片：5

较小的叶片数可以是的模型更能够捕捉到训练数据中的噪声，此处参数的设定是我们根据实际操作选择的最优化参数。

叶子节点最小样本数：8

这个值默认是 1，由于数据量巨大，一般我们采取的方式是增大这个值，以保证在大数据量计算的前提下可以做到一定的“剪枝”效果。

基于上述对于随机森林模型的了解，我们得到线下预测 RSRE 的均方误差为：9.584

7.3 模型比较

在 7.1 和 7.2 中所提及的模型的 RMSP 值都是线下进行计算，为了展现不同模型的泛化能力，弄够运用在特定场景或者未知场景下，我们将其进行线上评分，并将上述所有模型得到线上线下两部分 RMSP 进行直观对比，如下所示：

表 14.各模型 RMSP 值

模型	全连接神经网络	深度神经网络联合线性回归	随机森林
线下 RSRP 值的 RMSE	9.698	9.953	9.584
线上 RSRP 值的 RMSE	9.434	/	/

注意的是线下我们用神经网络模型相关模型时对数据集进行了 200 万训练集，100 万测试集的划分，在对随机森林进行数据集划分时，将 1100 万数据作为训练集，100 万作为测试集；线上我们均采用的是 1200 万训练集。

同时，因为最先运算的全连接神经网络在线上的得分达到 9.434，因此，我们以此为标准，若线下某模型的得分小于 9.434，我们有理由认为该模型可能会在线上表现更好，也就有理由认为该模型会有优于全连接神经网络的可能，若大于或者等于，说明该模型与前者表现较为一致，因此出于线上条件的限制，我们不再对其进行线上得分计算。

八．模型优缺点及展望

8.1 模型优点与缺点

8.1.1 优点

1.在设计整体特征体系时，本文通过阅读大量文献及相关资料，找到较多校正因子，并将之纳为考量范围之内，保障了特征体系的完整性的同时进行了方向性的考虑，由正反向思维逻辑共同思考完成了特征设计。
2.在后期特征筛选的过程中，基于已有特征体系下，通过数据预处理后通过是否可达，是否发散，是否相关的三步步骤进行一一分析，并且在计算相关性时，出于对数据本身属性考虑，将三种相关系数一一对比，一一排序，从而在宏观层面上达到逻辑化要求，微观层面上达到可视化要求，横向对比不同相关系数，列向对比不同特征排序，达到了多层面分析要求。
3.出于对题目的理解，本文主要以递进思维进行特征设计，环环相扣，层层递进，具有较强逻辑性。
4.在建立 AI 模型过程中，通过对多种模型的分析，对比选择最优模型的同时，就整体上而言均方根误差 RMSE 值均较小，一定程度上说明了整体模型建立的优越性。

8.1.2 缺点

1.在传播损耗中，未穿透损耗等带来的影响，此类因子通常在模型中表现为常数没，但在不同环境下，不同地形下，不同地物类型下的求得的常数是不同的，因此说明此类因子与地物类型等有关，可是出于计算复杂性，因此我们将之通过机器学习技术来研究，而非通过实测数据进行探究，是特征提取中的问题所在。
2.出于对于设计特征的不同理解，本文在问题一中设计的部分设计特征是在现有条件下无法进行计算的，所以在问题一中并未进行详细计算，只是对于部分校正特征进行计算，以此进行合理性分析，可能会与原始意图有一定出入。

8.2 展望

本文研究工作主要由选择了传统 cost231-hata 模型及奥村哈特模型为基础作为展开，通过其中涉及到的校正因子，参数等作为部分特征，以机器学习技术为手段建立 AI 智能模型，但是在无线传播模型这一领域中还有很多问题有待进一步研究与解决，同时在机器学习技术中也有待更好更适用的模型的建立，由于时间与条件限制，今后关于上述工作还有待进一步探究：

1.对于奥村，奥村哈特以及 cost231-hata，SPM 模型下涉及到的参数和校正因子并不是适用于所有未知情况的，尽管本文有对相应地形，传播环境进行考虑，但是由于数据的限制，类似与街道宽度等特征我们无法进行计算，也没有相应实测数据,因此若是有这些参数数据，怎样用好这些参数数据并运用到模型当中去还有待进一步研究

2.传统传播模型主要是用在宏蜂窝环境下，若是要针对于微蜂窝，微微蜂窝，例如室内预测模型的研究也是当前非常重要的研究课题。

3.基于射线跟踪的电磁理论的微蜂窝模型，虽然数据计算量非常庞大，但是的确对于实际研究的意义非常大，在未来相关研究中可能也可以利用机器学习技术进行对该模型的计算等，随着计算机技术的提高与发展，该方向可能也会成为未来主要研究方向之一。

九．参考文献

[1]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.

[2]付昆鹏，移动通信中电波传播模型的研究[D],云南师范大学，2011

[3]张永华，专网无线网络规划中的传播模型研究[D]，北京邮电大学，电子工程学院，2019.

[4]罗淑婉，杨庚. TD-SCDMA 无线传播模型校正[J]. 广东通信技术, 2007(1).

[5]刘晶. 基于 LTE 系统的传播模型及模型校正算法研究[D]. 东华大学, 2014.

[6]无线传播理论[Z]，华为技术有限公司，2001.

[7]马庆，王志鹏. 数据挖掘在无线网络规划中的应用研究[J]. 信息通信, 2016(8):262-263.

[8]晏远为，黄正彬，叶春显. 基于栅格化的 LTE 无线网络覆盖预测仿真[J]. 信息通信, 2015(11):200-202.

[9]邓中亮，肖占蒙，贾步云.城市空间无线定位信号传播模型校正方法研究[J]. 导航定位与授时, 2017(03):15-20.

[10]于仰源，孙宜军，王磊. 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).

[11]杨辉，朱曦宁，尧文彬. 基于地物类型矢量距离的 TD-LTE 无线传播模型映射方法[C], 中国移动通信集团设计院新技术论坛. 2014.