第十五届中国研究生电子设计竞赛

技术论文

中文题目：数据驱动的机器学习算法为基站链路损耗预测与站址规划赋能

英文题目：Data-driven machine learning algorithms empower base station link loss prediction and site planning

参赛单位：南京航空航天大学

队伍名称：继续奔跑的小白

指导老师：李海林

参赛队员：杨凌辉、杨昌林、张嘉纹

完成时间：2020-7-20

摘 要

在无线通信网络的规划与建设中，高效精确的无线链路损耗预测对于基站网络优化有着重要的指导意义。无线信道的电波传输特性与信道环境密切相 关。复杂的电磁通信环境下，会产生诸如反射、散射、绕射等电波传播方式，此时的链路损耗与自由空间损耗模型大有不同，需要加入特定的场景修正因子。精准的无线电传输损耗模型的建立，可以对基站目标通信覆盖范围内的电磁传输情况进行相对准确的预测，为后续通信业务指标例如：小区基站覆盖范围、小区间网络干扰、通信传输速率以及通信链路容量等指标的有效估算提供理论依据。传统电波传输损耗模型主要有三种，经验模型、理论模型和改进经验模型。经验模型的设计是根据大量场景实测数据进行公式参数拟合，典型模型有Cost231-Hata、Okumura等。理论模型主要依据电磁传播理论、考虑电磁波在空间中的反射、折射和散射，借助几何光学理论、几何绕射理论和一致绕射理论等，进行理论建模，代表性的是Volcano模型。改进经验模型是依据特定电波传输场景下的实测数据，设定特定的场景修正因子，从而获得更为细化的分类场景传输模型，典型的有Standard Propagation Model, SPM。近年来随着通信设备数量的增多，以及大数据云计算等存储能力和算力的应用和普及，产生了大量的基站信号测量数据，本文基于丰富的实测数据，提出了基于随机森林算法RF (Random Forests, RF)、CNN (Convolutional Neural Network, CNN) 和DNN (Deep Neural Network, DNN) 网络的链路损耗预测模型，与传统模型预测精度进行对比分析。并在此基础上，创新性地引入基于加权聚类算法的KW-means模型对基站的最优部署进行优化，结果表明，该算法可以有效减少基站弱信号覆盖比例，从而更好地服务基站用户。

关键词：链路预测，聚类算法，站址规划，RF，CNN，DNN

Abstract

In the planning and construction of wireless communication networks, efficient and accurate wireless link loss prediction has important guiding significance for the optimization of base station networks. The radio wave transmission characteristics of wireless channels are closely related to the channel environment. In a complex electromagnetic communication environment, there will be radio wave propagation methods such as reflection, scattering, and diffraction. At this time, the link loss is very different from the free space loss model, and a specific scene correction factor needs to be added. The establishment of an accurate radio transmission loss model can make a relatively accurate prediction of the electromagnetic transmission within the target communication coverage of the base station, for subsequent communication business indicators such as cell base station coverage, inter-cell network interference, communication transmission rate and communication chain The effective estimation of indicators such as road capacity provides a theoretical basis. There are three main models of traditional radio wave transmission loss, empirical model, theoretical model and improved empirical model. The design of the empirical model is based on a large number of scene measurement data for formula parameter fitting. Typical models include Cost231-Hata and Okumura. The theoretical model is mainly based on the theory of electromagnetic propagation, considering the reflection, refraction and scattering of electromagnetic waves in space. The theoretical modeling is carried out with the help of geometric optics theory, geometric diffraction theory and uniform diffraction theory. The representative is the Volcano model. The improved empirical model is based on the measured data in a specific radio wave transmission scenario, setting a specific scene correction factor, so as to obtain a more detailed classification scene transmission model, typical are Standard Propagation Model, SPM. In recent years, with the increase in the number of communication devices, and the application and popularization of storage capacity and computing power such as big data cloud computing, a large amount of base station signal measurement data has been generated. Based on rich measured data, this paper proposes network link loss prediction models based on random forest algorithm (Random Forests, RF), CNN (Convolutional Neural Network, CNN) and DNN (Deep Neural Network, DNN) ,they are compared with the prediction accuracy of traditional models. On this basis, the KW-means model based on the weighted clustering algorithm is innovatively introduced to optimize the optimal deployment of the base station. The results show that the algorithm can effectively reduce the weak signal coverage ratio of the base station to better serve the base station users.

**Keywords:** link prediction, clustering algorithm, site planning, RF, CNN, DNN

目 录

第1章 作品难点与创新 1

1.1 作品难点 1

1.2 作品创新点 2

第2章 方案论证与设计 3

2.1 基站链路损耗数据集 3

2.2 特征设计 6

第3章 原理分析 14

3.1 传统链路损耗模型 14

3.2 随机森林 16

3.2.1 随机森林回归模型 16

3.2.2模型参数分析 17

3.3 CNN网络 17

3.4 DNN网络 18

3.5 加权KW-Menas算法 19

第4章 软件设计与流程 21

4.1 随机森林模型 21

4.2 CNN模型 21

4.3 DNN模型 22

第5章 系统测试与误差分析 26

5.1 随机森林模型预测结果 26

5.1.1 预测数据统计 26

5.1.2 分布密度统计 27

5.2 CNN模型预测结果 28

5.2.1 预测数据统计 28

5.2.2 分布密度统计 29

5.3 DNN模型预测结果 29

5.3.1 预测数据统计 30

5.3.2 分布密度统计 31

第6章 客户端设计 32

6.1 系统架构图 32

6.2 数据可视化客户端 32

6.3 站址规划客户端 32

第7章 总结 36

参考文献 37

第1章 作品难点与创新

## 1.1 作品难点

精确的无线电波传输损耗模型的建立，可以对基站通信覆盖范围内的链路损耗情况进行相对准确的预测，为后续基站部署以及网络容量规划提供理论依据。合理的基站站址选择也是影响覆盖范围内网络通信质量的关键因素之一，稀疏的基站部署会导致覆盖区域信号质量差、通信容量不足等问题，过密的基站分布又会引入较大的站间同频信号干扰，同时也会增加运营商基建成本和运维难度[1]。因此，准确的链路预测以及合理的基站部署是无线网络规划中重要的影响因素。但是，传统的移动电波通信的基站部署问题属于NP-hard问题[2]，结合实际部署需求引入的参量以及约束条件会对模型的求解带来维度灾难，不能给出高效、精确的部署方案。因此，为了避免传统模型在站址规划问题上的弊端，需要提出一种合理有效的通用解决策略。

现有无线传播模型主要有经验模型、理论模型和改进型经验模型[3]。经验模型通过统计分析大量数据得到无线电波的路径损耗公式，主要有Cost231-Hata模型、Okumura-Hata模型等。改进型经验模型则在传统经验模型的基础上引入了更多的计算因子，从而适应更多的场景，如标准传播模型（Standard Propagation Model，SPM）。这些模型计算简单，易于部署，因此被大量地在工程运用，但是其精度较低，无法做出精确的预测。理论模型则根据电磁波理论，计算发射机与接收机之间的所有信号传播路径，又称为多射线模型，如Volcano模型。这种模型预测准确，但运算的复杂度较高且需要精确的电子地图。目前经验模型及其改进型更为常用，但不同场景中传播模型的准确度不同，需要校准模型参数提高传播损耗预测的准确性。

在基站站址规划方面，国内外研究主要集中在不同基站规划约束条件下的模型算法改进。文献[4]基于基站的覆盖范围和通信容量约束条件，结合元启发式算法，求解该优化问题，问题的求解需要诸多先验参数，例如基站机械参数、发射参数以及所处地理位置参数等，因此模型复杂度较高，很难优化。文献[5]针对4G蜂窝网提出混合整数模型对网络基站位置进行最优迭代规划。文献认为基站的空间分布结构对最佳站址选择起着重要的决定作用，并采用三种空间随机模型，结合公开数据库数据进行站址选择仿真，以最大化基站覆盖准确度为衡量依据对不同模型效果进行对比分析。文献[6]以最大化基站下行平均遍历容量为目标，基于粒子群算法对基站天线端口位置进行优化。

上述基站网络规划研究主要根据不同规划场景约束条件，对模型进行最优化求解，且传统模型在输入参数较大的情形下算法收敛很慢甚至不收敛，同时传统算法以大量假设为前提，简化模型的数学分析，在实际基站部署过程中，具有一定的局限性。

## 1.2 作品创新点

近年来，人工智能与机器学习技术在大数据的驱动下，从上世纪的理论研究走向了实际应用场景，在很多领域取得了成功。机器学习技术也随着计算架构的进步，其低复杂度和高实时性使得其与无线通信的结合成为了可能。传播损耗的预测也可以看作利用频率、高度、距离、发射机、接收机、建筑物等输入特征组成的特征向量，寻找输入特征向量与目标值传播损耗间关系的回归问题。机器学习是解决回归问题的有效工具，可以高效地用于传播损耗模型预测。

针对传统无线信号传播模型存在的不足，该文基于4000个基站的不同场景实测数据，引入了基于随机森林、CNN、DNN算法的无线传播损耗预测模型。站址规划方面，在基于机器学习的链路预测基础上，考虑测试点与基站的空间相对位置，在原有聚类KMeans分簇算法基础上引入信号覆盖权重，提出了基于数据挖掘的KW-Means分簇算法，该算法将权重映射到测试点与基站相对的水平和垂直方向上，计算基站的二维偏移平均量，确定迭代位置，对基站部署进行优化。最终根据基站覆盖阈值以及目标函数的迭代更新，得到当前模式下最终的基站拓扑。通过大量的实测数据仿真结果表明，随机森林与KW-Means算法相结合的网络优化系统，极大提升了基站对弱信号点的覆盖比例，提升了网络通信质量。

第2章 方案论证与设计

## 2.1 基站链路损耗数据集

实验数据集由华为公司2019年全国研究生数学建模A题提供，包含4000个基站小区与下属测试点，其中包括小区站点与接收机所在栅格的地理及物理信息信息。地理数据被划分为五米见方的栅格，并用栅格的坐标来表示，同一个栅格内的物体位置统一为栅格西北方向角坐标。数据集数据格式及含义：

表2-1数据集数据向量含义及其格式

|  |  |  |
| --- | --- | --- |
| 字段名称 | 含义 | 单位 |
| Cell Index | 小区唯一标志 | - |
| 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 Building Height | 小区站点所在栅格建筑物高度 | m |
| Cell Altitude | 小区站点所在栅格(Cell X,Cell Y)海拔高度 | m |
| Cell Clutter Index | 小区站点所在栅格(Cell X,Cell Y)地物类型索引 | - |
| X | 栅格位置，X坐标 | - |
| Y | 栅格位置，Y坐标 | - |
| Building Height | 栅格(X,Y)上建筑物高度 | 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 | dBm |

下图显示了数据集中各个主要特征的分布情况，其中传播损耗近似正态分布在110dB附近。基站天线与终端之间的水平距离集中在1000米以内，基站天线高度则基本集中在24米，对于传统模型而言可以基本覆盖。基站天线的俯仰角多下倾八度，基站天线方位角则多由60度、180度、300度这三个角度组成，每个天线负责120度的蜂窝区域。基站发射机功率为多个定值。海拔高度也基本呈在500米为均值的正态分布。建筑物则主要集中于底层，且范围很大，这个在后面对特征的箱形图分析中也可以看到。

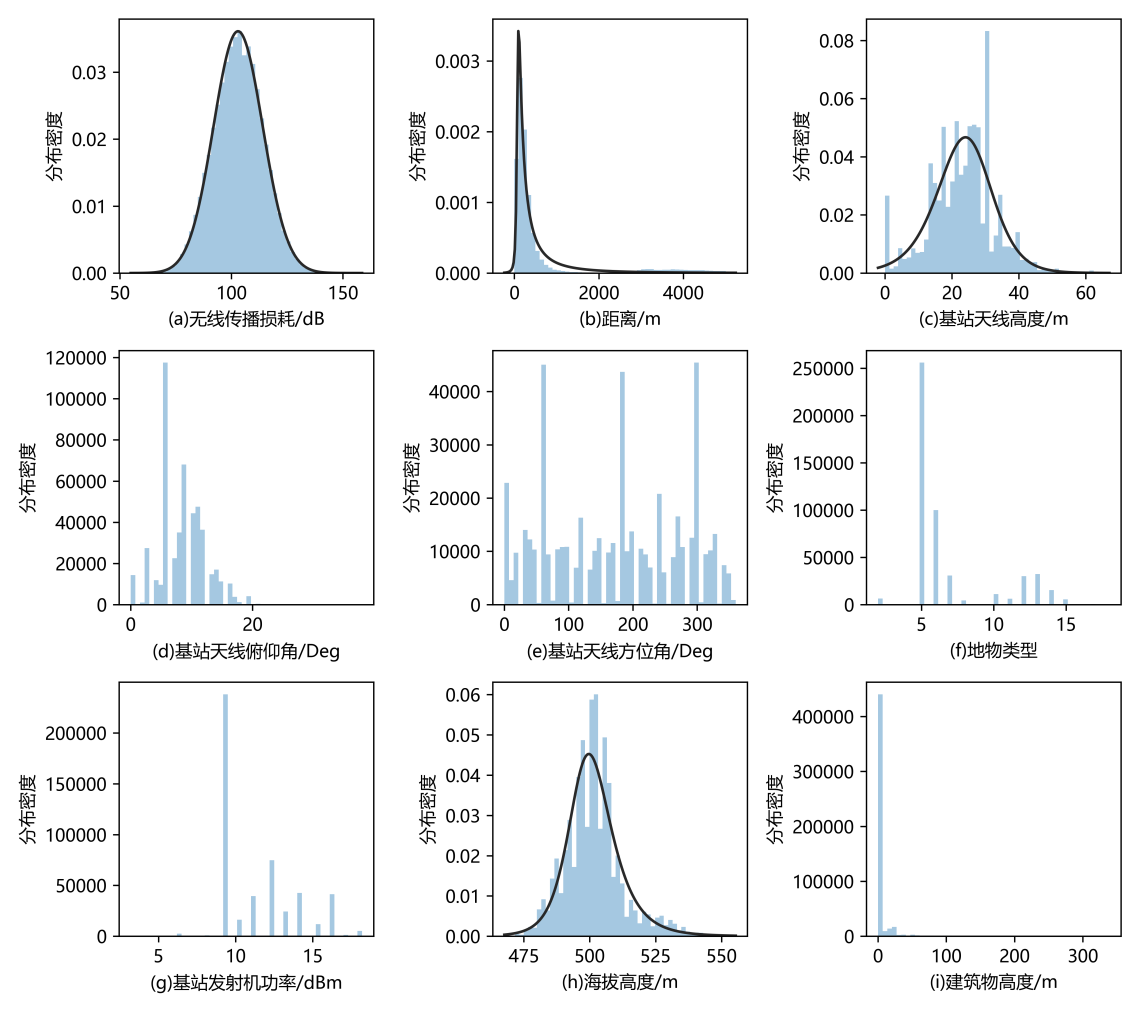


图2-1数据集特征分布情况

放大地图中心区域，可以看到水系、建筑区和树木的分布，其中空白区域为无记录区域。

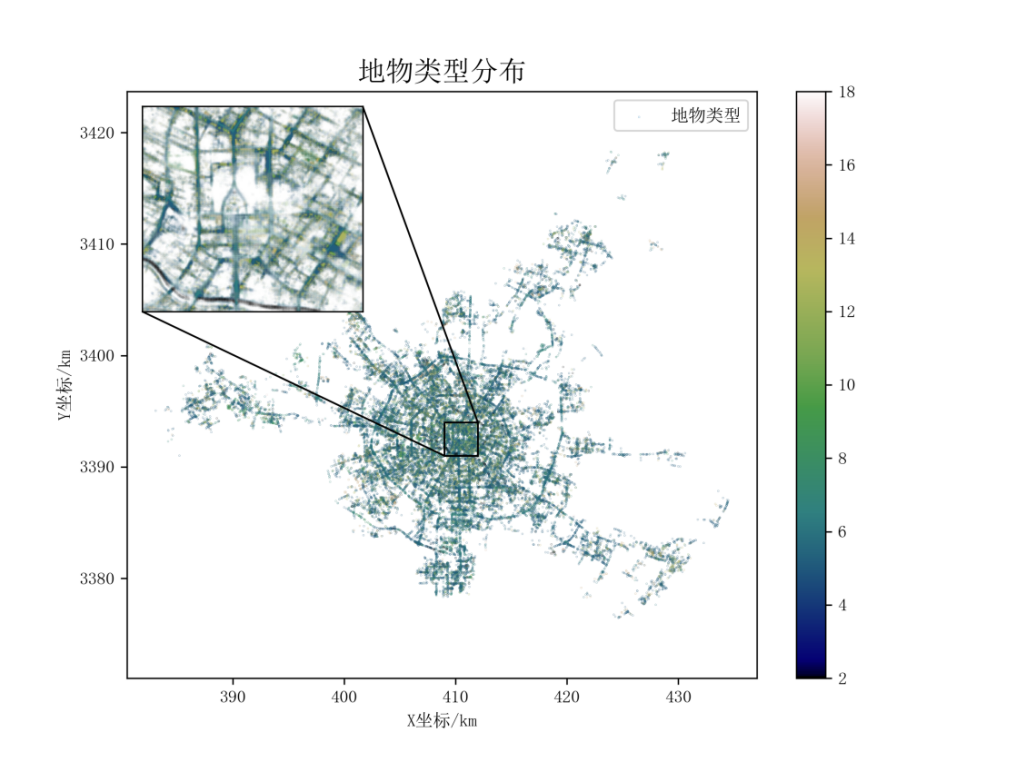


图2-2电子地图中地物分布

接下来对城市的海拔高度分布进行统计，如下图，城市内部的海拔净差约为100米，西北方向和东南方向的山区间分布着主要城区，因此在特征设计的时候在涉及高度信息时，需要考虑海拔因素。

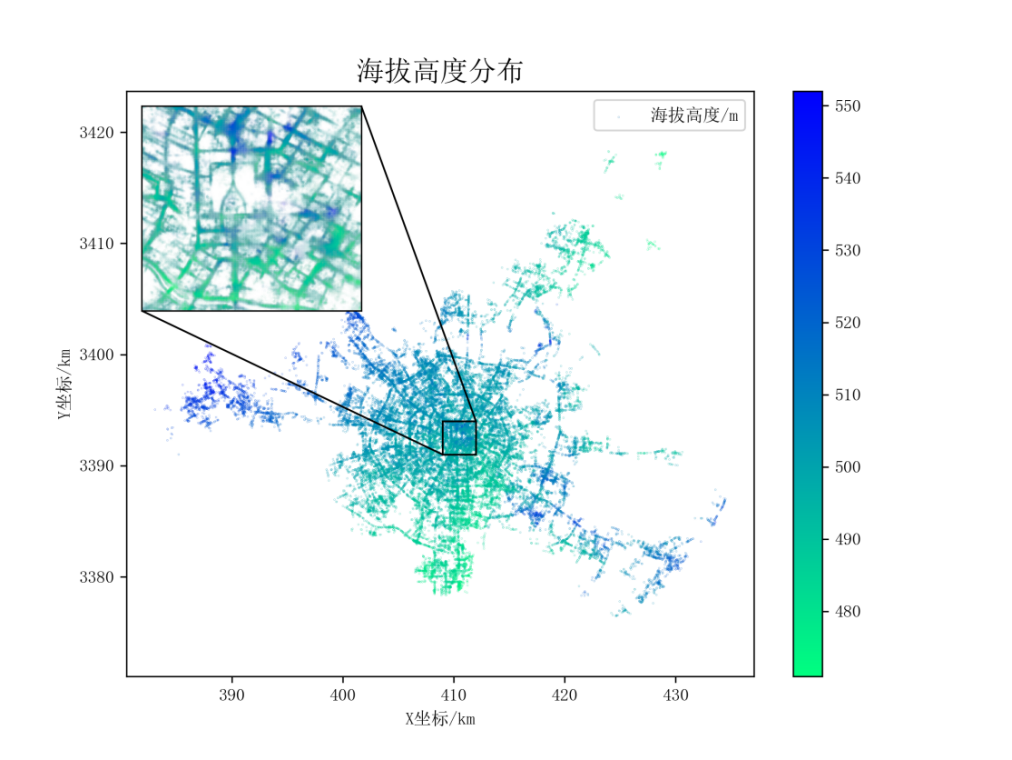


图2-3电子地图中海拔高度分布

下图对城市建筑物高度进行统计，城市建筑最高高度为400米以上，其中底层建筑物分布最为广泛。

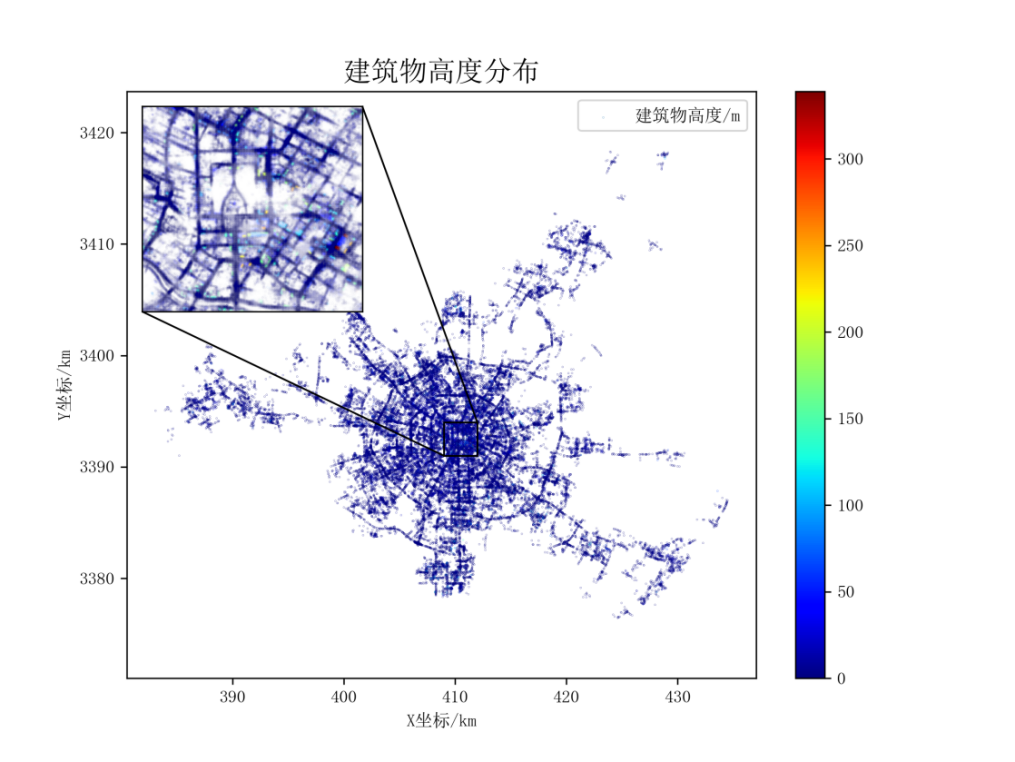


图2-4电子地图中建筑物高度分布

下图为城市中心区域的信号强度分布，其中红色标点标志发射机，可以看到街道、河流、树木等相对空旷的区域平均信号接收功率更高，这也与无线传播理论相符，在这些地理环境中，电磁波的传播可以近似为含地面反射的双射线模型，且遮挡造成的衰减较小。同时，街道两侧的建筑物也会对电磁波进行反射，形成类似于波导的结构，也利于电磁波的传播。

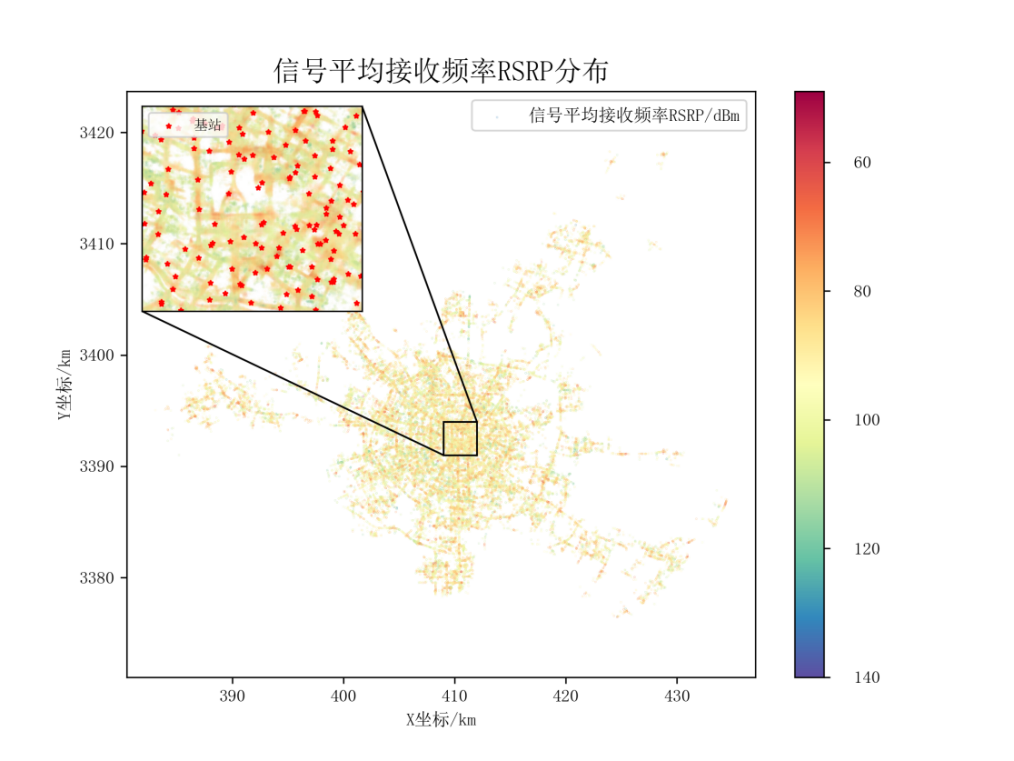


图2-5基站及信号接收功率分布

## 2.2 特征设计

机器学习算法主要依赖输入变量与输出变量之间的关系进行学习与预测，因此输入变量也被成为特征，而一组输入变量被称为特征向量，输出变量也被成为目标。特征向量与目标都是服从一定分布的随机向量，机器学习算法试图从他们之间的统计意义出发描绘特征向量与目标之间的关系。特征工程可以寻找到与目标值之间相关性更高、包含信息更多的特征，使得机器学习算法对变量之间的高阶关系有更精准的预测，同时还可以减少运算开销。

特征设计主要通过理论分析中已知对目标值路径损耗的影响因素与对数据集中数据特点得到。通过理论知识将原始数据集中的特征转化，使新的特征具有物理意义或统计意义，对于地理坐标或者设备编号等没有绝对物理含义或者人为设定的原始数据，在预测未知区域与未知设备时则会失去意义，进而干扰预测的准确性，则应将其转化为具有普遍物理含义的特征或者去除，主要包括栅格坐标与小区编号。同时需要设计新的带有绝对物理含义的特征，保留原始数据集的信息。

无线信道环境复杂多变，除了电磁波在不同介质和环境中传播带来的损耗外，收发天线增益以及天线方位角等也会对接收端信号接收功率造成影响。由此可见基站下行链路测试点信号质量指标主要和基站端、传输环境以及站间信号干扰叠加等因素有关。

除了电磁波在不同介质和不同环境中传播带来的传播损耗外，收发天线的增益也影响了收发功率的衰减量。同时，由于天线的方向性导致不同方向上的天线增益不同，因此发射机和接收机之间的俯仰角和方位角同样会对接收端信号功率衰减产生影响。下面主要对上述影响因素进行分析，设计新的特征。

本文通过对数据集进行分析并参考传统模型，以此为基础进行特征设计。

2.2.1特征设计

**1) 距离与高度**

利用数据集中水平坐标位置，由此得到基站栅格与接收机栅格水平距离；同时考虑海拔高度因素的基站栅格与接收机栅格三维距离；同时考虑海拔高度和基站高度的基站天线与接收机栅格三位距离。其中分别为终端与基站的位置坐标。

 （2-1）

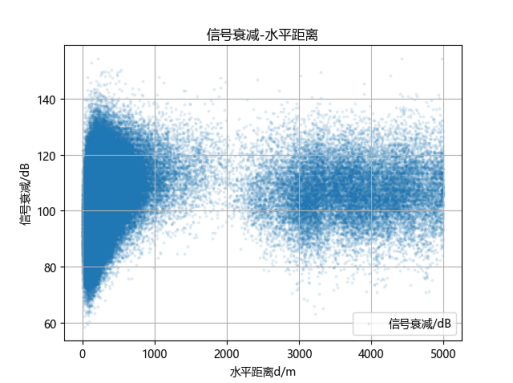


图2-6基站到接收机水平距离与信号衰减关系

上图为全部样本中距离与信号衰减的关系，随机取样二十万点。距离在两千米以内，信号的衰减与距离正相关，但是仍然有一个不确定的区间；在两千米之外，衰减分散在一定范围之内，且均值较为确定。三种距离与衰减的关系相近，图像无法显示其间区别，应用数值型指标对其进行分析选择。其中分别为终端与基站的海拔，为基站高度。

 （2-2）

 （2-3）

在传统经验模型中，发射机和接收机的高度是无线路径衰减的重要影响因素，数据集中包含基站高度信息，可以用作参考。下图中基站高度明显会对信号衰减范围影响，因此应当作为特征之一。栅格与信号线相对高度如下图所示，为天线主轴沿线与栅格的相对高度。信号线高度主要与天线方向性增益、发射机高度、绕射等特征有关的综合性特征。其中分别为基站天线的机械与电下倾角。

 （2-4）

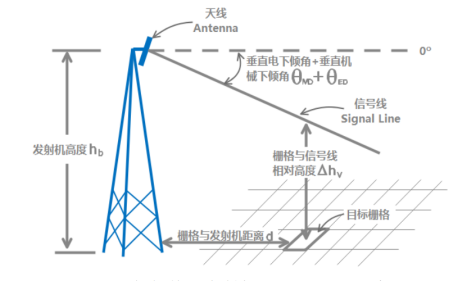


图2-7信号线高度示意图

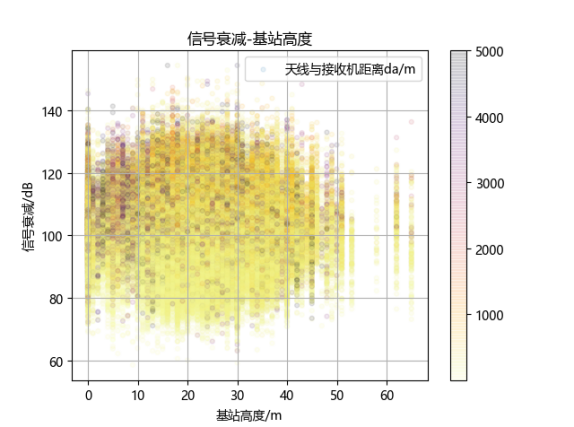


图2-8基站高度与信号衰减关系

**2)方向性**

由于天线方向性的影响，因此应当设置接收器栅格与天线主轴的方位角夹角、俯仰角夹角作为特征。而方位角夹角与俯仰角夹角因地物在不同角度上不均匀的分布，导致其也是一个综合性特征。方位角夹角同时与地物分布有关；由于天线均下倾，因此俯仰角夹角也与水平距离有关。

 （2-5）

 （2-6）



图2-9天线主轴垂直夹角与信号衰减关系



图2-10天线主轴水平夹角与信号衰减关系

从图2-10中可以看到随着与天线主轴水平夹角的增大，信号衰减的范围有了明显的变化，同时颜色深度代表的距离分布也较为均匀，说明方位角夹角特征适用于各个距离的接收机。

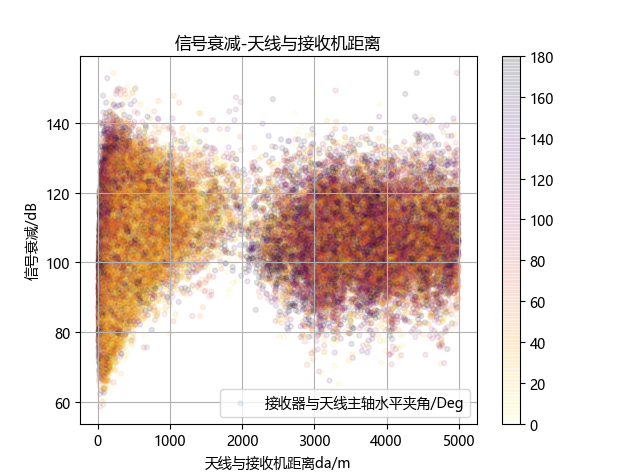


图2-11基站天线到接收机距离与信号衰减关系

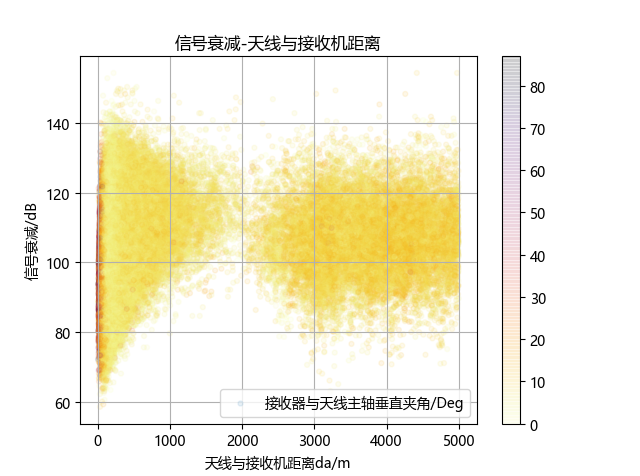


图2-12基站天线到接收机距离与信号衰减关系

在原本距离与信号衰减的图像中加入方位角夹角和俯仰角夹角进行联合分析，可以发现在两千米内不同的角度信息下，信号衰减与距离的关系呈现不同的正相关关系，因此在不同的方向下，天线的增益与地物不同的分布对距离与信号衰减的关系产生了影响。而对于两千米以外的接收机栅格，不同的方向分布较为均匀，因此天线方向性在远端影响较小。

**3)地物类型**

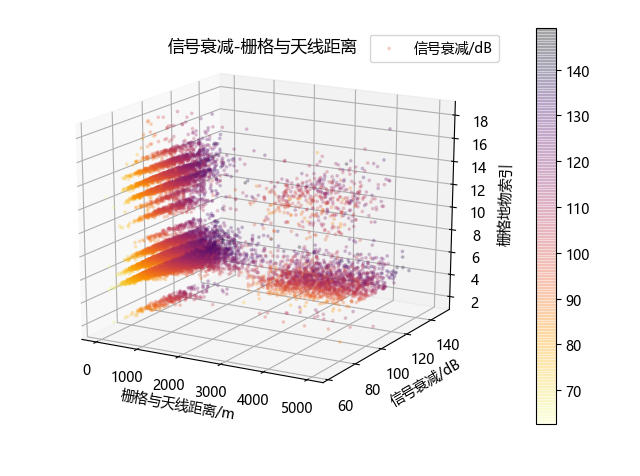


图2-13不同地物类型下基站天线与接收机距离与信号衰减关系

不同的地物类型下，无线信号会进行不同路径的传播，因此其传播特性会有所改变。同样使用距离与信号衰减的关系对不同地物类型进行观察。可以看到在不同的地物类型下，信号衰减与距离的关系有较大的差异，因此应当作为特征考虑。

**2.2.2特征分析**

通过数据分析与特征设计，提出了五个新的特征，其分布如下图。对于一数据向量，偏度表征统计分布的偏斜方向与程度：

 （2-7）

峰度表征统计分布的陡缓程度的统计量：

 （2-8）

其中正态分布的偏度与峰度为0。如下图所示，大部分特征的峰度与偏度在零附近，接近正态分布。其中发射机中心频率偏度6.70、峰度43.94，建筑物高度偏度6.06、峰度52.52，信号线高度偏度-3.92、峰度20.56，三个距离特征偏度2.65、峰度5.70，基站天线主轴垂直夹角偏度-2.56、峰度11.24，以上特征分布集中且不均衡。

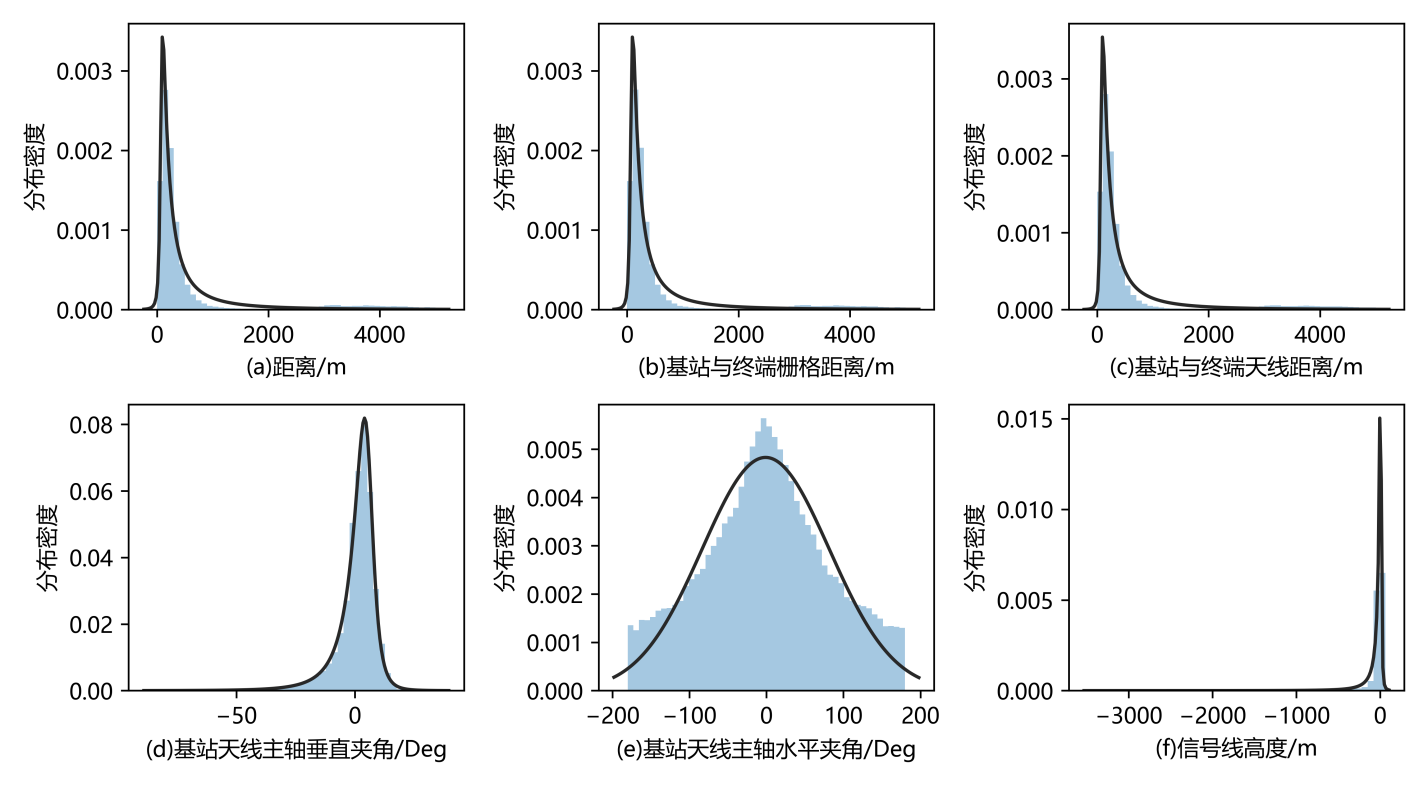


图2-14设计特征分布

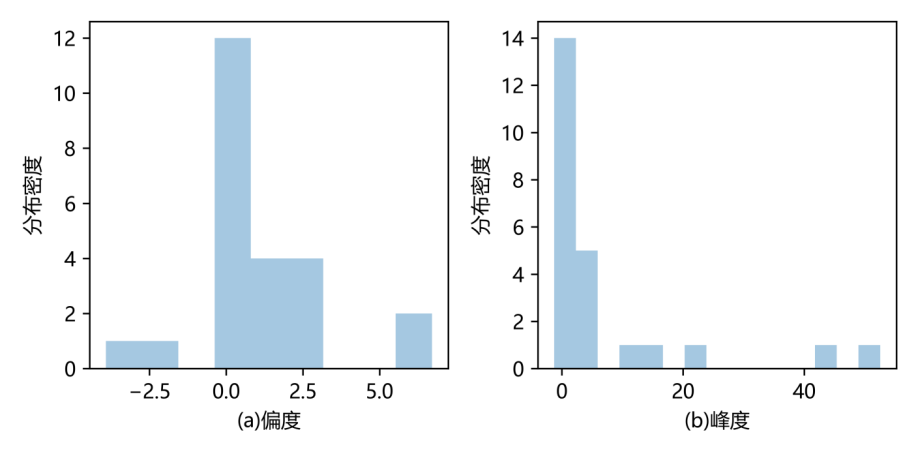


图2-15特征分布指标

**2.2.3训练特征集**

表2-2特征集含义及单位

|  |  |  |
| --- | --- | --- |
| 特征名称 | 含义 | 单位 |
| Distance | 基站与栅格的水平距离 | m |
| Signal Line Height | 栅格上小区发射机信号线高度 | m |
| 3Axis Distance | 小区站点与栅格距离 | m |
| Antenna Distance | 小区发射机与栅格距离 | m |
| Antenna Downtilt | 小区发射机天线主轴与栅格方向垂直夹角 | Deg |
| Antenna Azimuth | 小区发射机天线主轴与栅格方向水平夹角 | Deg |
| Cell Index | 小区唯一标志 | - |
| 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 Altitude | 小区站点海拔 | m |
| Cell Building Height | 小区站点建筑物高度 | m |
| Cell Clutter Index | 小区站点地物类型索引 | - |
| X | 栅格位置，X坐标 | - |
| Y | 栅格位置，Y坐标 | - |
| Altitude | 栅格海拔 | m |
| Building Height | 栅格建筑物高度 | m |
| Clutter Index | 栅格地物类型索引 | - |

# 第3章 原理分析

## 3.1 传统链路损耗模型

3.1.1 COST231-Hata模型

COST231-Hata 模型是由Hata模型发展而来的，其频率适用范围扩展到2GHz以上，基站覆盖范围大于一千米，终端天线的高度在1-10米，基站天线高度在30-200米。其路径损耗的计算经验公式为：

 （3-1）

其中传播路径损耗单位为dB，载波频率单位为MHz，基站天线有效高度单位为米，用户天线高度纠正项，基站到终端距离单位为千米以及场景纠正常数。其中为用户天线高度。

 （3-2）

是城市校正因子。

 （3-3）

为小区场景校正因子，

 （3-4）

3.1.2SUI模型

SUI（Stanford University Interim）模型是由802.16 IEEE组织和斯坦福大学联合推出的适用于城市郊区环境的信道模型，其主要的适用范围为传播距离，接收天线的高度，基站天线高度。模型将地物类型分A为有着高密度树木覆盖的丘陵地区，传播衰减最高；B为树木稀少的丘陵地区，传播衰减居中；C为少量树木覆盖的平地，传播衰减最低。其传播损耗计算公式为：

 （3-5）

其中传播距离的限制条件为，单位为米。为其它介质与杂波造成的阴影衰落。其中：

 （3-6）

 （3-7）

波长单位为米。基站天线有效高度单位为米。如下表所示，由地形决定：

表3-1 SUI模型参数设置

|  |  |  |  |
| --- | --- | --- | --- |
| 参数 | 地形A | 地形B | 地形C |
| a | 4.6 | 4.0 | 3.6 |
| b | 0.0075 | 0.0065 | 0.005 |
| c | 12.6 | 17.1 | 20 |

模型的频率与终端天线高度的修正因子为：

 （3-8）

 （3-9）

其中频率的单位为MHz。为终端天线的高度，单位为米。

3.1.3ECC-33模型

ECC-33模型为基于Okumura模型的扩展模型，其主要适用范围为信号中心频率附近。不同于Okumura模型在东京城郊获取的实验数据，ECC-33模型则使用欧洲中小城市的实验数据。路径损耗公式为：

 （3-10）

其中分别为自由空间传播损耗，基本中位路径损耗，基站高度影响因子和终端高度影响因子。其定义如下：

 （3-11）

 （3-12）

 （3-13）

对于中小城市：

 （3-14）

其中基站天线高度、终端天线高度单位为米。传播距离单位为千米。频率单位为GHz。

## 3.2 随机森林

### 3.2.1 随机森林回归模型

回归问题主要研究一个随机变量与多个变量组成的随机向量分布之间的关系。其中对于在某一范围内的值，目标取值可以由一定的概率分布描述。由于为随机变量，因此回归函数对其的预测为其根据概率分布函数求得的数学期望。回归函数也就是描述随机变量分布期望值与随机向量分布的函数[7]。

随机森林回归模型与随机森林分类模型相对，不使用类别目标而是使用数值作为预测值。随机森林的组成取决于不同但服从同一分布的随机向量。假设训练集的每次采样是从随机向量中独立进行的，则数值预测变量的均方泛化误差为：

 （3-15）

随机森林是由将K个决策树集合的预测结果进行平均进行预测，可得以下定理：

定理3.3：当随机森林中决策树的数目趋近于无穷时，可以概率1得

 （3-16）

将公式右侧使用泛化误差代替，则有定义森林中决策树的平均泛化误差为：

 （3-17）

定理3.4：若对于任意，都有，则有

 （3-18）

其中为余项与之间的加权相关系数，其中之间独立。证明：

 （3-19）

公式右侧即为协方差，可以被写为：

 （3-20）

其中，加权相关系数定义为：

 （3-21）

遂可证：

 （3-22）

随机森林可以将决策树的平均泛化误差降低到相关性常数倍以下，因此提升随机森林性能的方法就是降低决策树之间的相关性与提升决策树的预测精度。在决策树个数、分裂选取特征数以及bootstrap抽样法抽样数这三个参数外，也需要对决策树的生长进行限制，从而获得最佳泛化误差，主要从直接限制决策树生长深度、限制节点分裂所需最小样本数、叶子节点最小样本数进行调节。

### 3.2.2模型参数分析

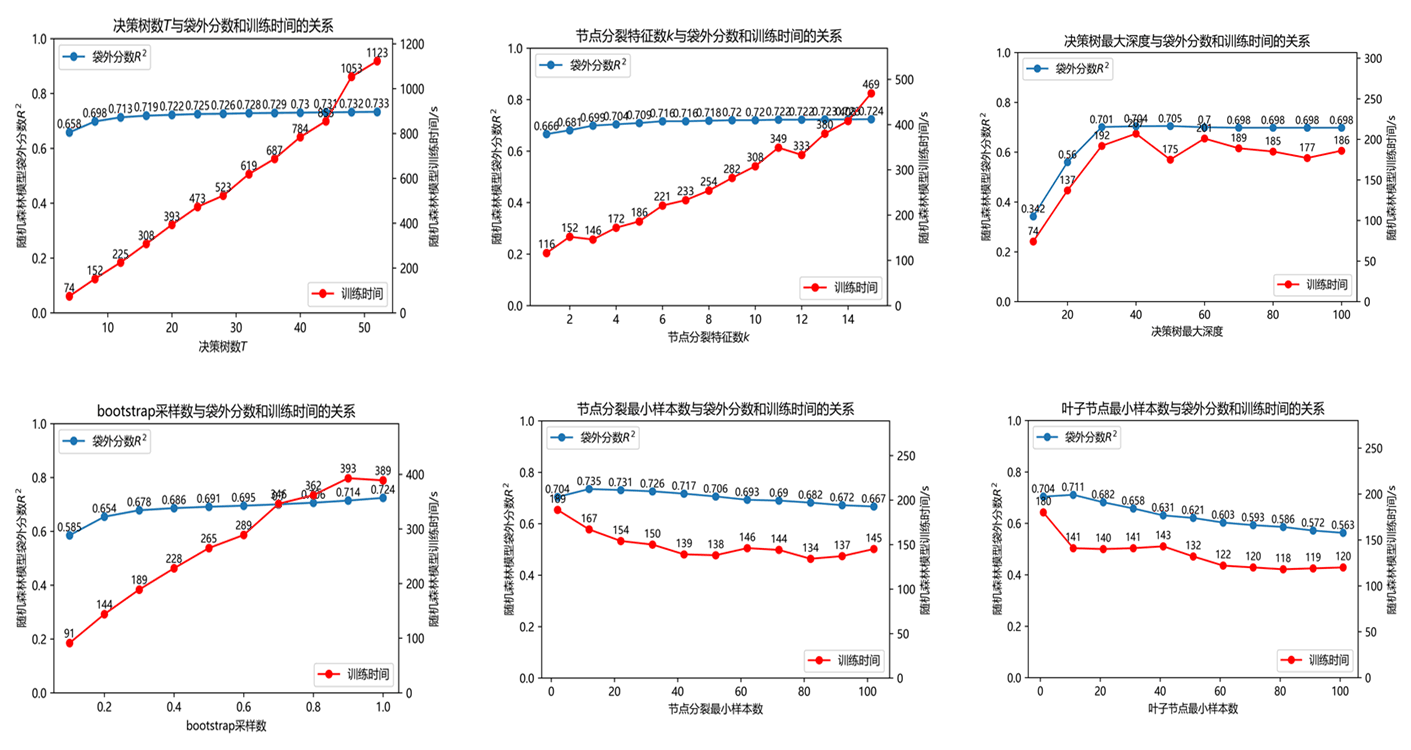
随机森林模型主要针对5个参数进行优化，决策树数n\_estimators，节点分裂特征数max\_features，决策树最大深度max\_depth，bootstrap采样法采样数max\_samples，节点分裂最小样本数min\_samples\_split，叶子节点最小样本数min\_samples\_leaf。其中bootstrap采样法采样数max\_samples采用比例形式，即真实采样数为数据集样本点数与max\_samples的乘积。

图 3-1 随机森林不同参数下的训练性能分析

## 3.3 CNN网络

将多个神经元模型按一定的层次结构连接起来，就能得到神经网络的模型。事实上，从计算机学科角度来看，我们可以不考虑神经网络是否真的模拟了生物神经网络，只需将一个神经网络模型看成一个包含了许多超参数的数学模型，这个模型有若干个激活函数组成。卷积神经网络主要包括输入向量的卷积层，池化层、全连接层以及神经元的激活函数[8]。

CNN的基本结构包括两层，其一为特征提取层，每个神经元的输入与前一层的局部接受域相连，并提取该局部的特征。一旦该局部特征被提取后，它与其它特征间的位置关系也随之确定下来；其二是特征映射层，网络的每个计算层由多个特征映射组成，每个特征映射是一个平面，平面上所有神经元的权值相等。

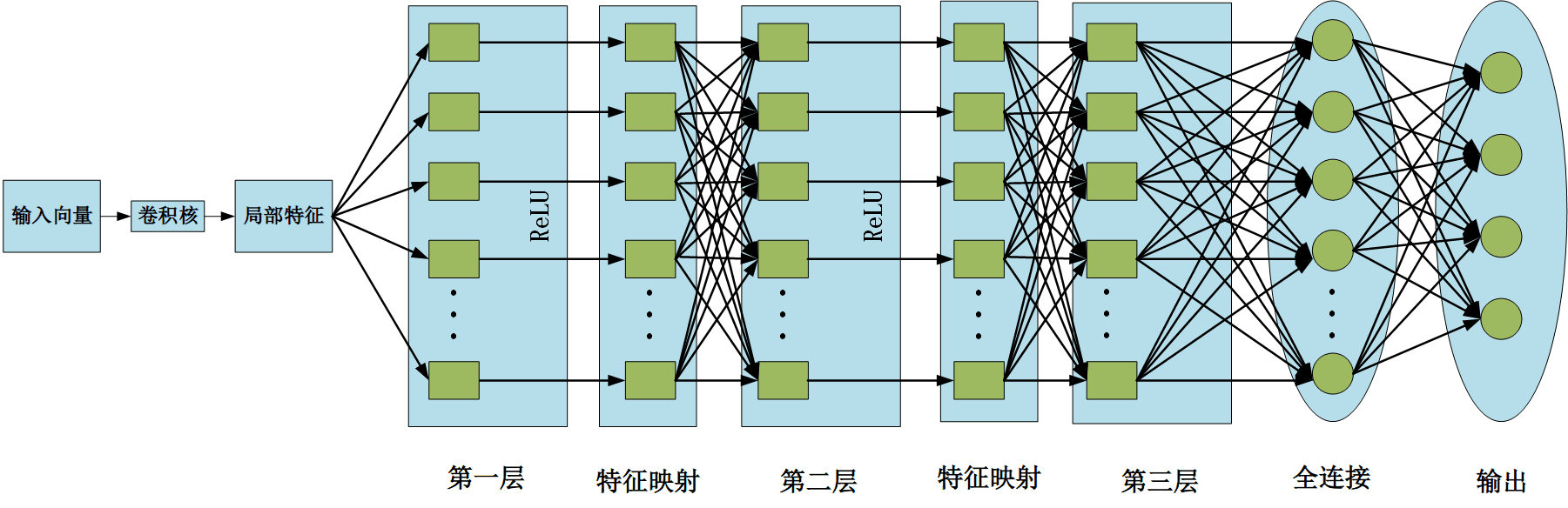


图 3-2 CNN深度神经网络原理图

假设输入向量为，卷积的离散形式为，其中为各层权重：

 （3-23）

二维卷积则表示为：

 （3-24）

## 3.4 DNN网络

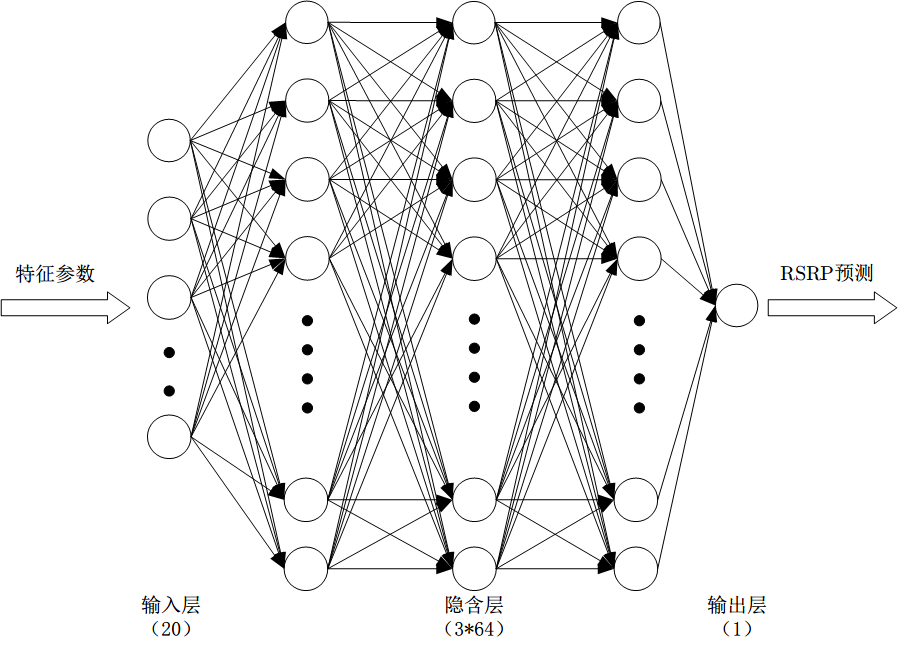
全连接深度神经网络DNN不同于CNN网络对向量进行卷积提取局部特征，其通过不同个数的神经元以及训练层次的深度来寻找输入特征向量和输出向量的映射关系。

图 3-3 DNN深度神经网络原理图

DNN网络模型训练过程存在梯度爆炸和梯度消失问题，‘Relu’激活函数的引入可以有效解决这两个困扰DNN的关键问题，

 （3-25）

第个隐藏层的输出为：

 （3-26）

其中，为第层第个节点与第层第个节点之间的权重。为第层第个节点的输出值。输出层输出为，其中为隐藏层节点数，代表线性输出，作为回归预测模型的输出层激活函数：

 （3-27）

## 3.5 加权KW-Menas算法

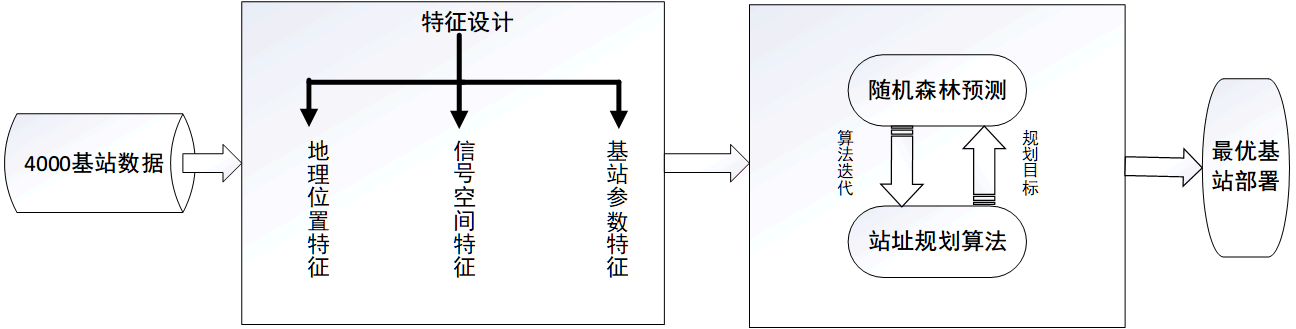
 本文提出了一种基于数据挖掘聚类算法的基站网络规划方案如图3-4所示。链路预测模块采用随机森林模型对接收点RSRP进行回归预测。输入变量与预测指标之间的相关性对预测精度起着决定性的作用，通过对原始数据特征进行分析，结合地理空间位置以及无线信道特性设计相关性较强的新的候选变量作为随机森林算法的输入，进行训练。聚类算法根据当前网络覆盖情况对基站位置进行重新调整，由于基站位置的变化，随机森林模型输入的特征向量需要重新计算调整，并预测得到当前基站分布模式下的RSRP指标。整个基站网络规划的目标函数为最小化低于覆盖阈值测试点RSRP的误差平方和，逐步减少基站下行链路中弱信号点的比例，保证网络中弱信号点的通信质量。

图 3-4 基站网络规划系统框图

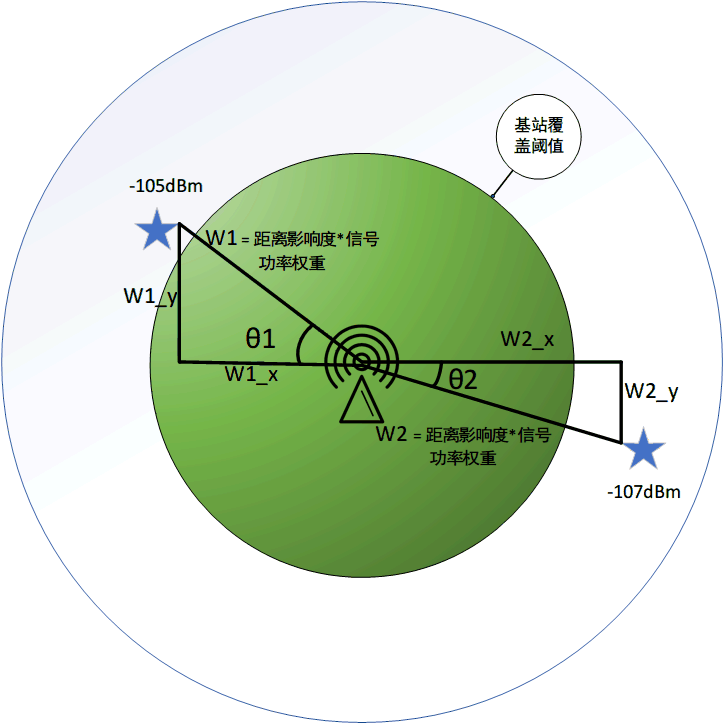
系统规划过程中，簇中心位置的调整根据距离影响度和信号覆盖权重来进行迭代更新。距离影响度基于测试点与簇中心的直线距离，信号覆盖较弱的测试点

图 3-5 基于二维坐标偏移的KW-means算法

对基站位置的影响应该较大，才能在逐步迭代优化过程中减少弱信号点的比例。

该文基于二维坐标偏移的KW-means算法，将距离影响度与覆盖权重的乘积投影到测试点与基站相对位置的二维坐标分布中来，具体如图3-5所示，求出簇集合中所有数据点在二维坐标中的总偏移量从而确定最新基站选址。

基站最新位置的迭代计算如下：

 （3-28）

 （3-29）

 （3-30）

 （3-31）

 （3-32）

当基站位置不再变化或者系统满足总的网络规划目标时，系统输出当前最优的基站部署位置。网络规划系统算法如下所示。



图 3-6 KW-menas站址规划算法

# 第4章 软件设计与流程

根据以上原理分析，我们设计了代码实现部分。硬件参数如下：CPU： i7-9750、显卡：GTX1650、内存：8GB。

## 4.1 随机森林模型

表4-1 随机森林模型参数设置

|  |  |
| --- | --- |
| 参数 | 数值 |
| 决策树数 | 20 |
| 节点分裂特征数 | 4 |
| 决策树最大深度 | 55 |
| 节点分裂最小样本数 | 7 |
| 叶子节点最小样本数 | 1 |
| Bootstrap采样数 | 9609466 |



图4-1 随机森林训练模型

## 4.2 CNN模型

表4-2 CNN模型参数设置

|  |  |
| --- | --- |
| 参数 | 数值 |
| 输入层数 | 4 |
| 隐藏层 | 3\*64 |
| 输出层 | 1 |
| 激活函数 | Relu |
| Dropout | 0.2 |
| 学习率 | 0.001 |
| 损失函数 | MAE |
| 训练Epochs | 100 |
| 优化函数 | RMSprop |
| 卷积核长度 | 2 |
| 训练百分比 | 80% |
| 交叉验证百分比 | 10% |
| 测试百分比 | 10% |

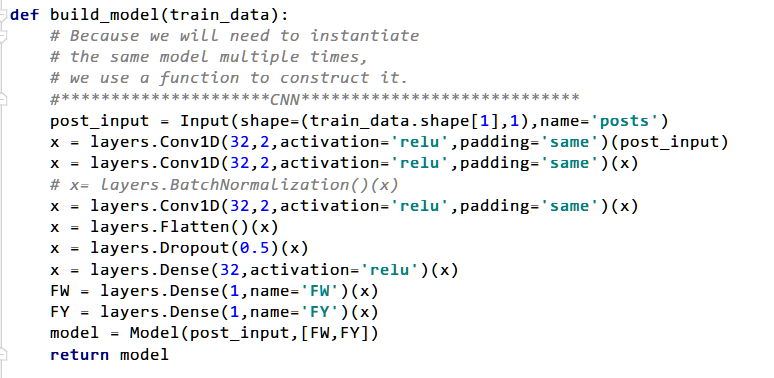


图4-2 CNN模型训练模块

## 4.3 DNN模型

表4-3 DNN模型参数设置

|  |  |
| --- | --- |
| 参数 | 数值 |
| 输入层数 | 4 |
| 隐藏层 | 3\*64 |
| 输出层 | 1 |
| 激活函数 | Relu |
| Dropout | 0.2 |
| 学习率 | 0.001 |
| 损失函数 | MAE |
| 训练Epochs | 100 |
| 优化函数 | RMSprop |
| 训练百分比 | 80% |
| 交叉验证百分比 | 10% |
| 测试百分比 | 10% |

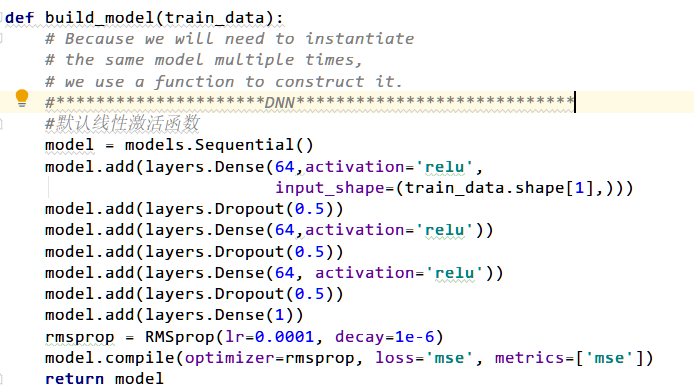


图4-3 DNN模型训练模块

## 4.4 KW-means



图4-4 KW-means核心算法



图4-5 KW-means迭代算法

图4-6 站址规划客户端主界面代码

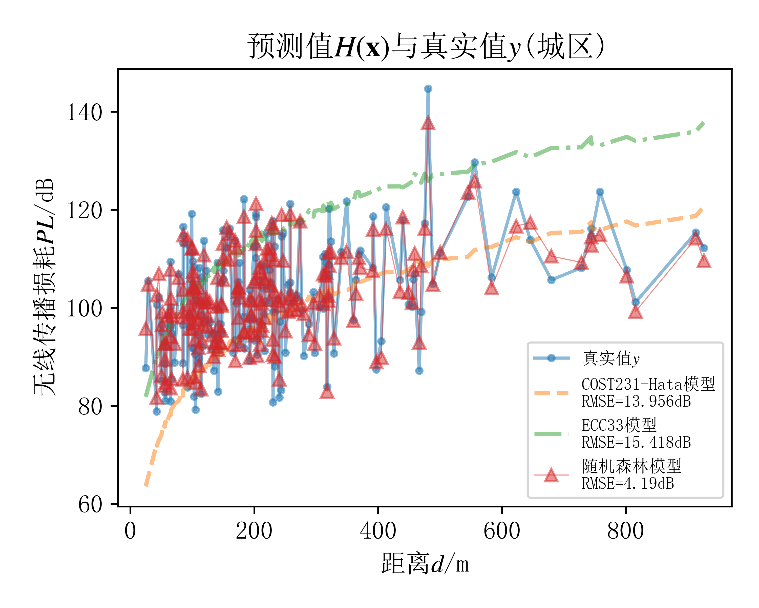


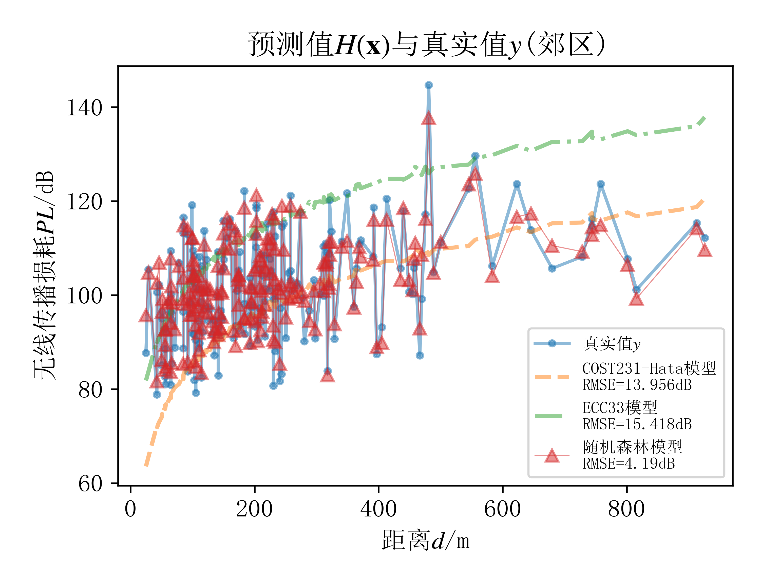
图4-7 数据可视化客户端主界面代码

# 第5章 系统测试与误差分析

## 5.1 随机森林模型预测结果

### 5.1.1 预测数据统计





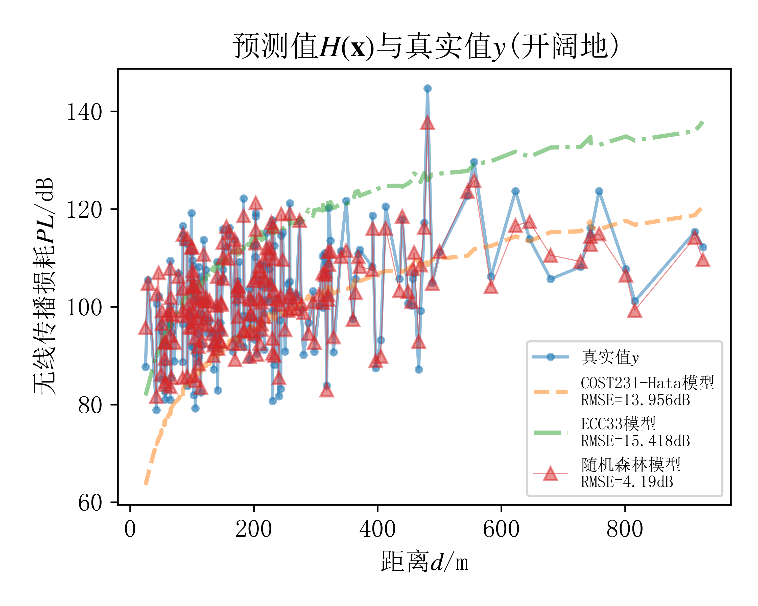


图 5-1 随机森林预测结果

随机森林由于可以对数据集进行较为精准的分割学习，因此其预测精度较传统模型更高。

### 5.1.2 分布密度统计

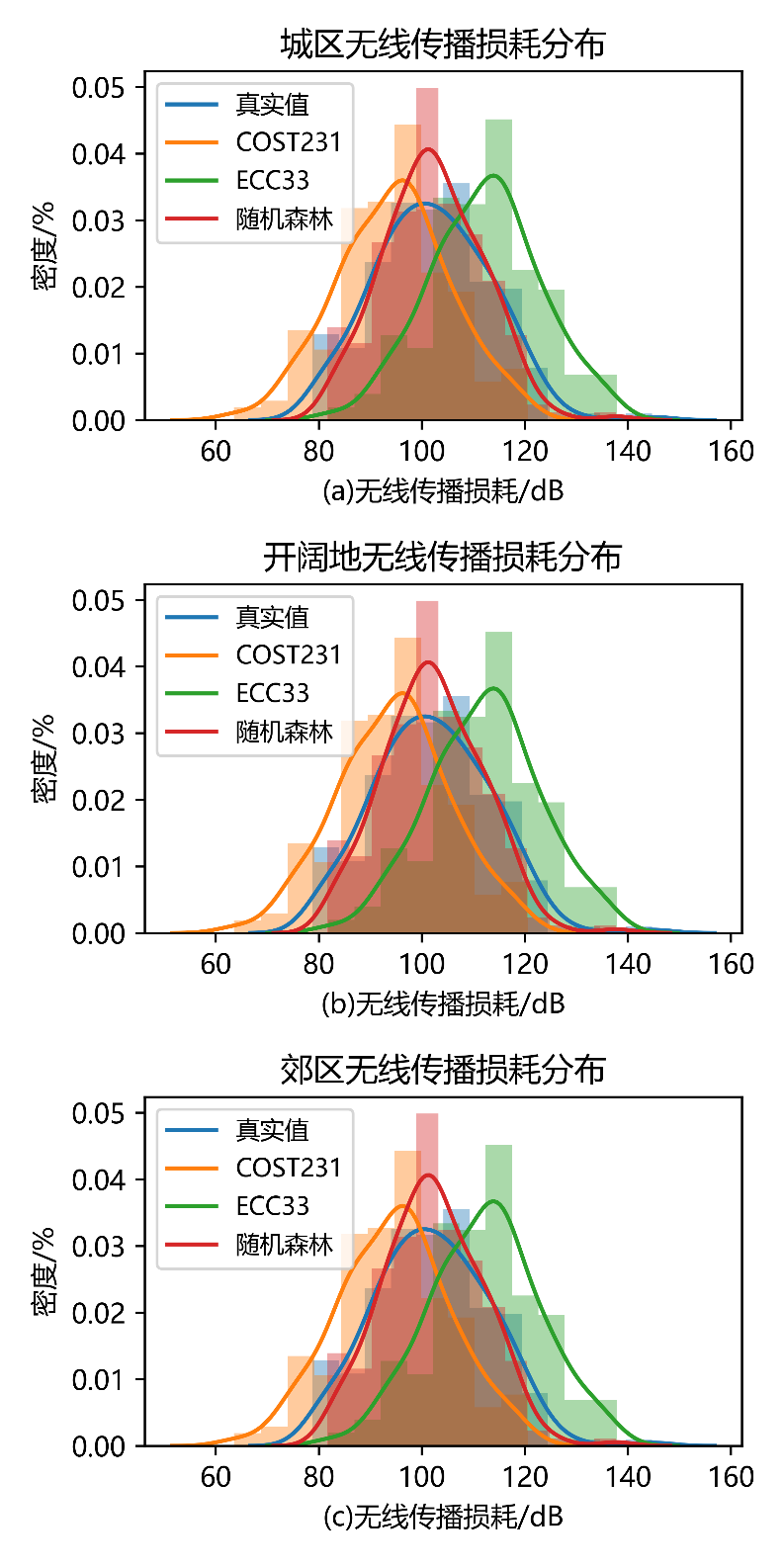
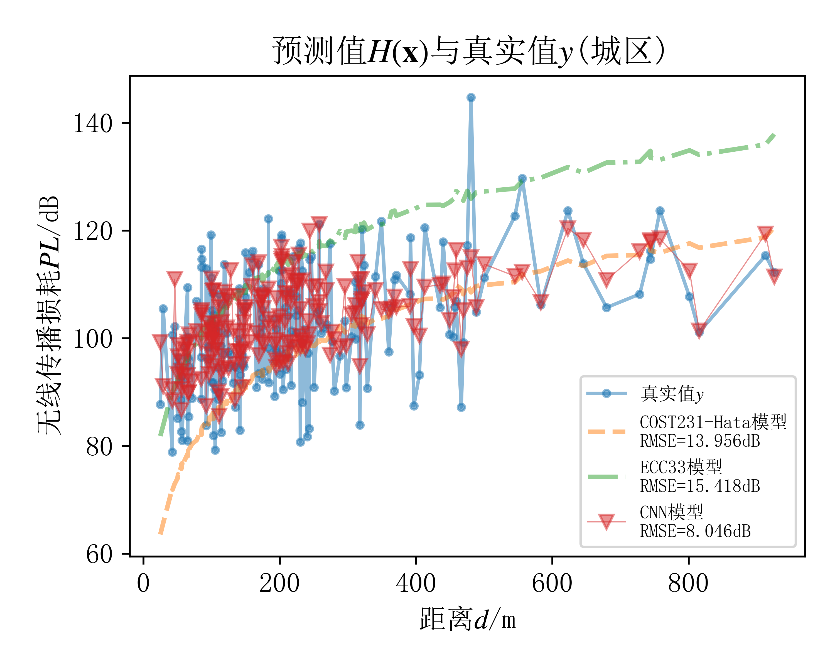


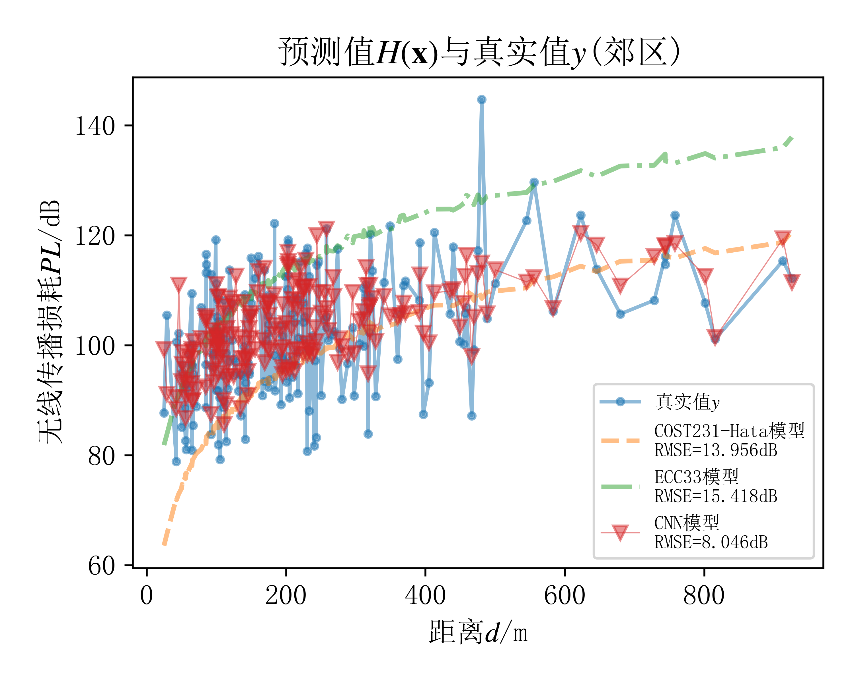
图 5-2 不同场景随机森林预测概率密度

最终预测结果的分布上，随机森林模型预测值与真实值之间的分布较为相近，而传统模型的估计结果与真实值之间有均值上的偏差。而随机森林模型预测值分布更加集中，则是由于其无法从训练集中提取出对小尺度衰落进行预测所需信息，因此无法对随机的小尺度衰落进行预测。

## 5.2 CNN模型预测结果

### 5.2.1 预测数据统计





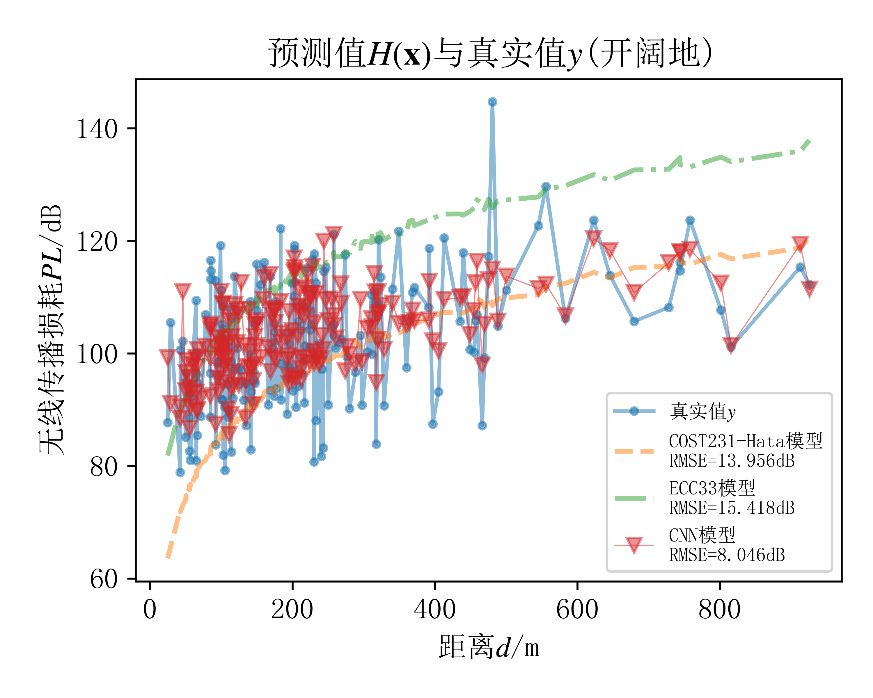


图 5-3 CNN模型预测结果

CNN对数据集进行较为精准的学习，因此其预测精度较传统模型更高。但受限于模型的大小与深度，精度较随机森林模型更低。

### 5.2.2 分布密度统计

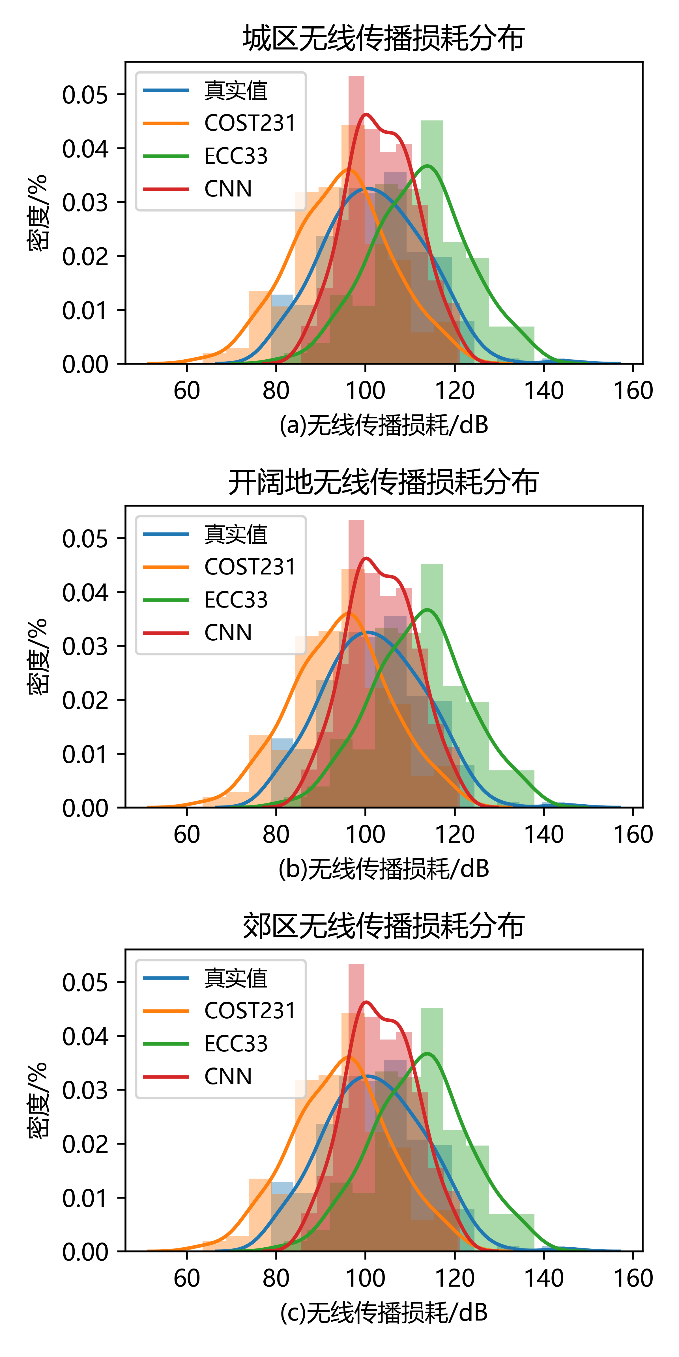
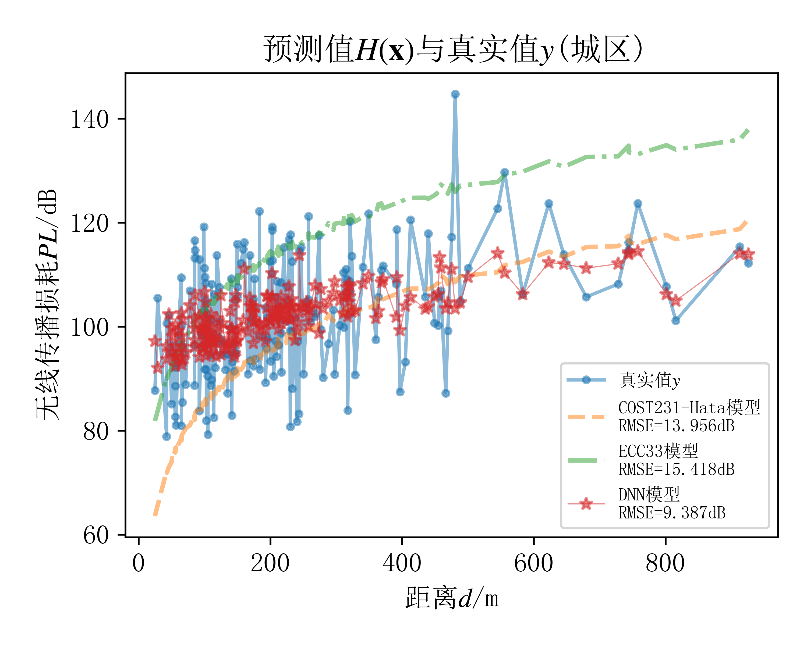


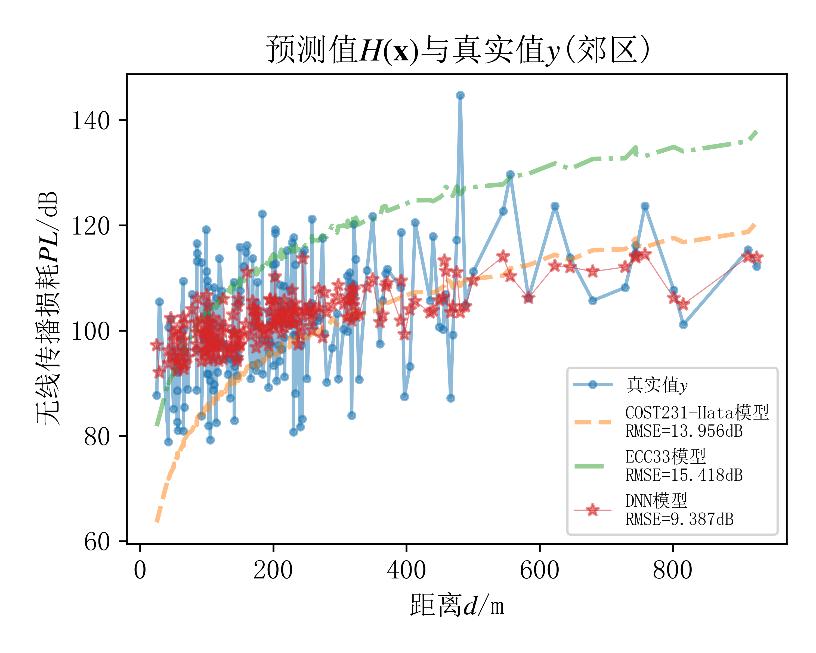
图5-4 CNN模型预测概率密度

最终预测结果的分布上，CNN模型预测值与真实值之间的分布较为相近，而传统模型的估计结果与真实值之间有均值上的偏差。而CNN模型预测值相较于随机森林模型分布更加集中，则是受限于其模型的大小。

## 5.3 DNN模型预测结果

### 5.3.1 预测数据统计





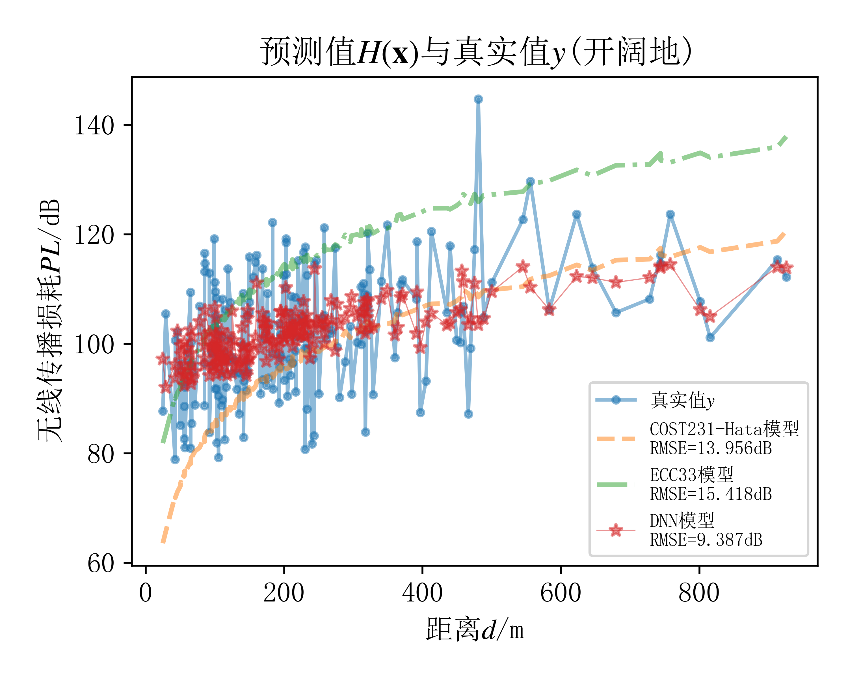


图 5-5 DNN模型预测结果

DNN对数据集进行较为精准的学习，因此其预测精度较传统模型更高。但受限于模型的大小与深度，精度较随机森林模型更低。

### 5.3.2 分布密度统计

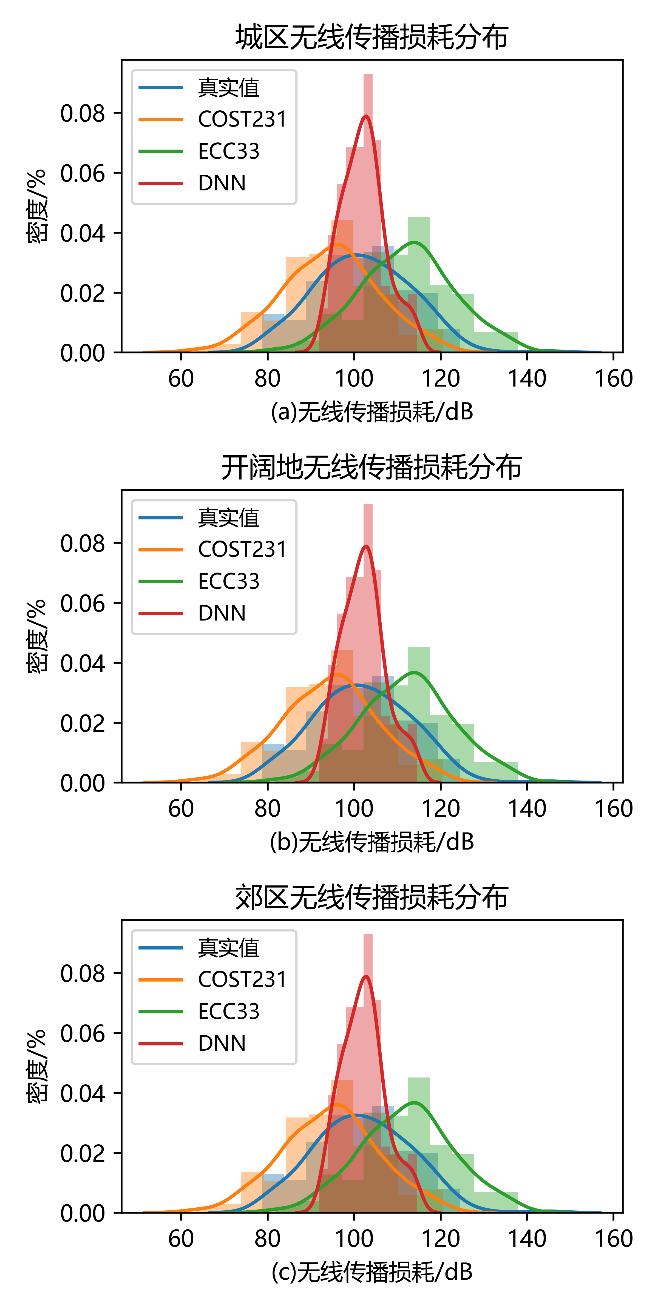


图 5-6 DNN模型预测概率密度

最终预测结果的分布上，DNN模型预测值与真实值之间的分布较为相近，而传统模型的估计结果与真实值之间有均值上的偏差。而DNN模型预测值分布类似于CNN模型，受限于其模型的大小，相较于随机森林模型分布更加集中。

# 第6章 客户端设计

## 6.1 系统架构图

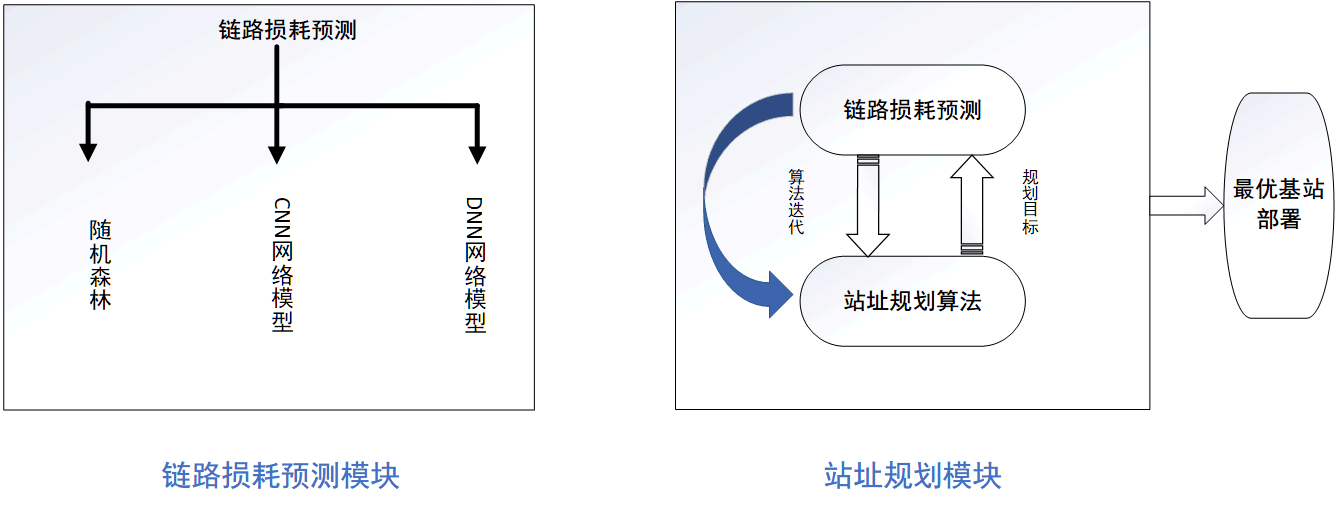


图 6-1 系统设计流程图

客户端后台设计流程图如上所示，客户端设计分为两个部分，1：基站链路损耗预测模块2：站址规划模块。预测模块根据上述章节基于选定特征向量的预测模型为基础，对测试数据进行预测，站址规划模块基于链路损耗预测结果以及站址规划目标进行最优迭代，从而确定站址最优部署。

## 6.2 数据可视化客户端

开发工具：Pycharm

开发语言：Python

编译器： python3.7

测试平台：win10

数据可视化客户端展示界面：

界面与用户的交互区域主要有1：选择水平观察距离范围2：选择垂直观察距离范围

界面展示区主要包括，1：观察范围内的地物类型索引可视化 2：观察范围内的海拔高度可视化。具体如下图所示。

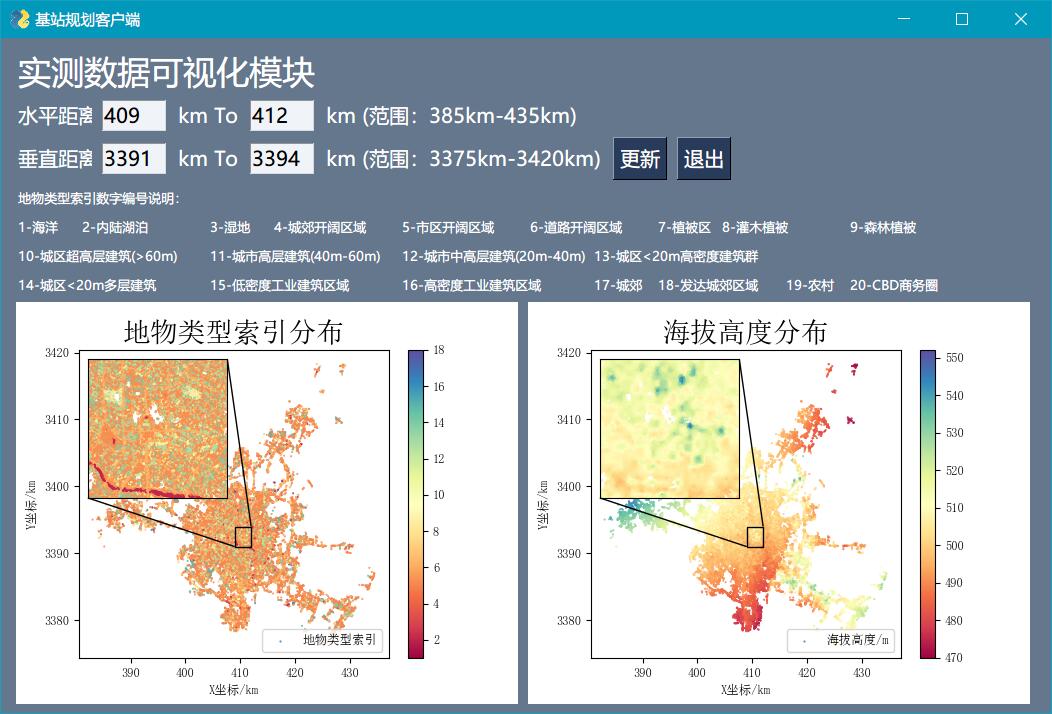


图 6-2 实测数据可视化模块

## 6.3 站址规划客户端

基站规划界面展示：

界面与用户的交互区域主要有1：链路损耗预测模型的选择2：基站规划的目标阈值选择。

界面主要包括四个画图区域和一个结果统计区域，左上绘制区域表示所选站址及其下属测试点的可视化数据，右上绘制区域表示站址规划前的基站下属测试点的接收信号能量分布，左下为所选模型的链路损耗预测精度与传统模型的对比，右下主要展示站址规划后的基站下属测试点的能量分布。站址规划算法可以在给定的基站阈值下，对站址位置进行最优迭代，从而减少弱信号（低于设定阈值）的覆盖比例，客户端顶部有站址规划前后的弱信号点数据的统计，可以看出KW-means的聚类算法在站址规划方向的可信性，在信息数据急剧爆炸的今天，为站址规划提出新的解决方案。

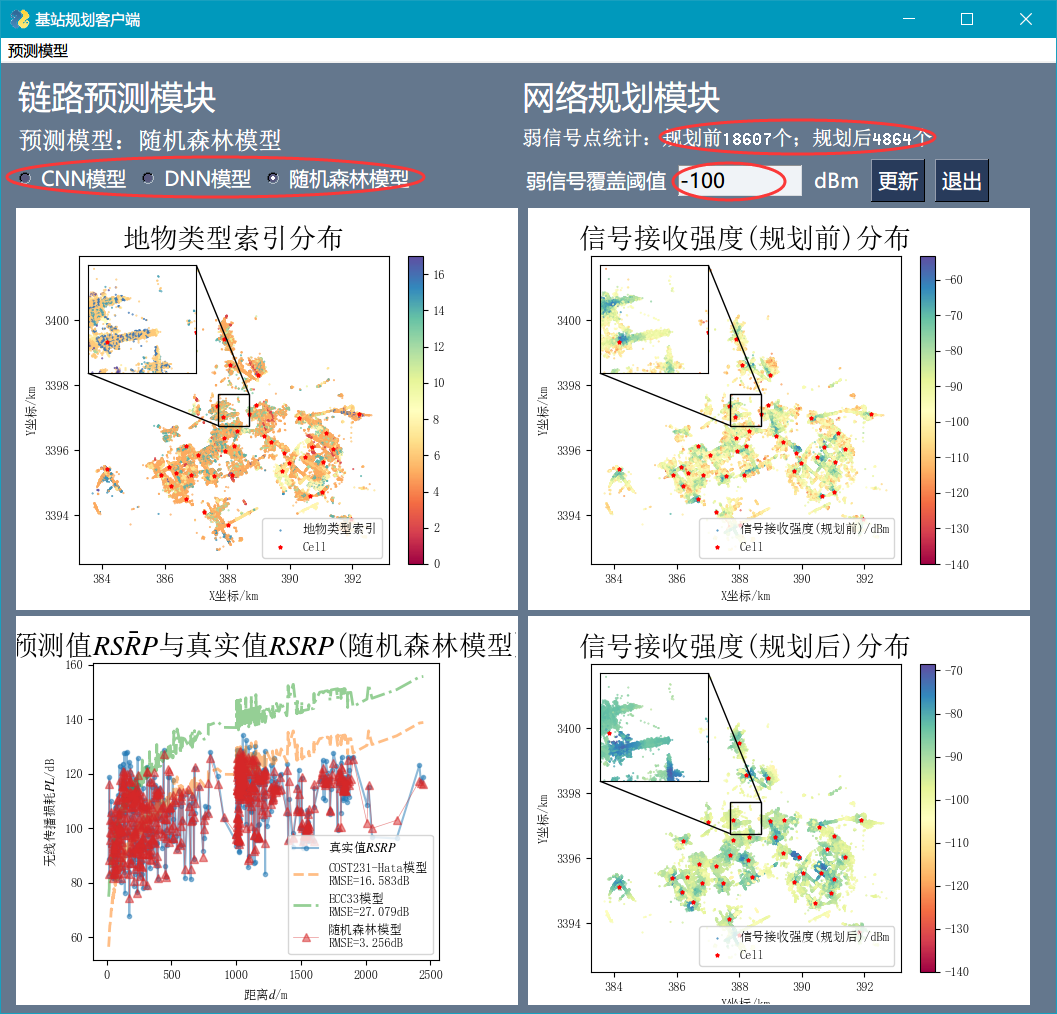


图6-3 站址规划客户端界面展示（-100dBm）

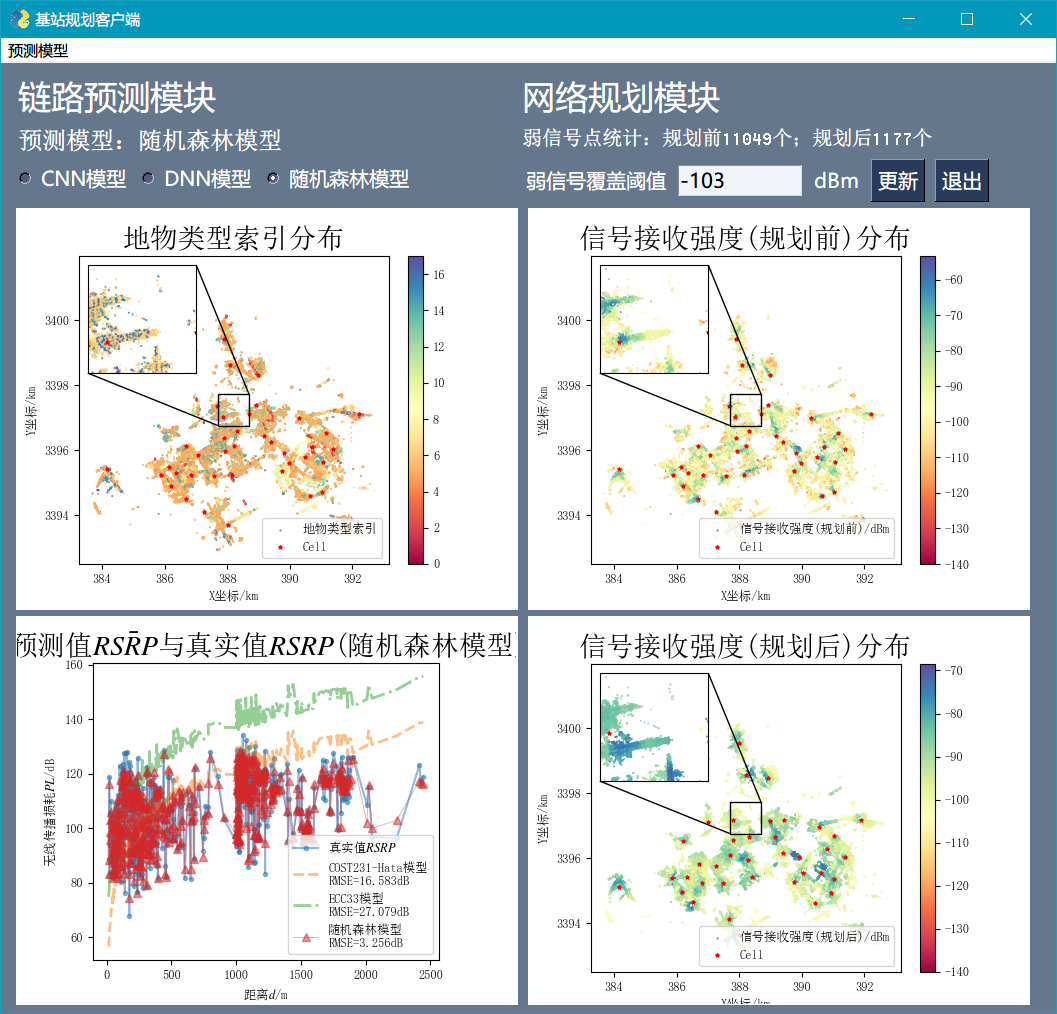


图6-4站址规划客户端界面展示（-103dBm）

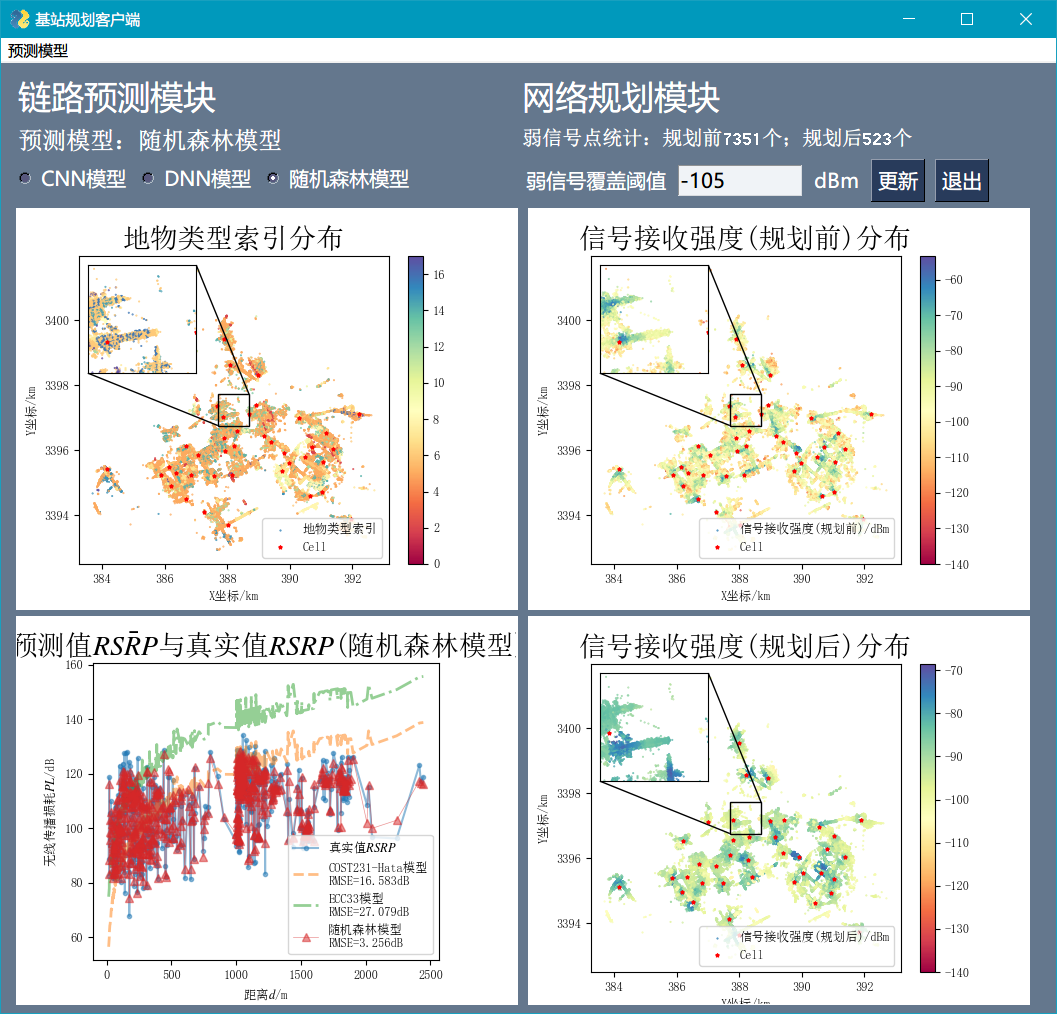


图6-5站址规划客户端界面展示（-105dBm）

# 第7章 总结

本文结合不同地域下的4000基站实测数据，提出了基于机器学习算法的链路损耗预测和站址规划算法研究。分别对随机森林算法、CNN神经网络和DNN神经网络进行了不同的仿真对比，仿真结果表明，数据驱动的基于机器学习算法的模型对不同场景下的损耗预测更加准确，这得益于机器学习算法对高维数据映射关系的较高提取能力。同时本文基于传统聚类KMeans算法提出了基于信号覆盖阈值和站址空间位置的加权KW-means算法对站址选择进行最优迭代。传统基站网络规划研究主要根据不同规划场景约束条件，对模型进行最优化求解，且传统模型在输入参数较大的情形下算法收敛很慢甚至不收敛，属于NP-hard问题，且模型以大量假设为前提，简化模型的数学分析，在实际基站部署过程中，具有一定的局限性。该文提出的KW-means算法将权重映射到测试点与基站相对的水平和垂直方向上，计算基站的二维偏移平均量，确定迭代位置，对基站部署进行优化。最终根据基站覆盖阈值以及目标函数的迭代更新，得到当前模式下最优的基站位置。通过大量的实测数据仿真结果表明，随机森林与KW-Means算法相结合的网络优化系统，极大提升了基站对弱信号点的覆盖比例，提升了网络通信质量。本文的研究对即将到来的5G基站大规模部署问题提出了新的解决思路，数据驱动的预测和规划模型可以有效地减少运营商在部署基站上的运营成本，同时可以有效保证信号覆盖区的用户体验，具有很好的理论参考价值。

# 参考文献

1. W. Zhao, S. Wang, C. Wang and X. Wu, "Approximation Algorithms for Cell Planning in Heterogeneous Networks," in IEEE Transactions on Vehicular Technology, vol. 66, no. 2, pp. 1561-1572, Feb. 2017, doi: 10.1109/TVT.2016.2552487.
2. E. Amaldi, A. Capone and F. Malucelli, "Planning UMTS base station location: optimization models with power control and algorithms," in IEEE Transactions on Wireless Communications, vol. 2, no. 5, pp. 939-952, Sept. 2003, doi: 10.1109/TWC.2003.817438.
3. Phillips C, Sicker D, Grunwald D. A survey of wireless path loss prediction and coverage mapping methods[J]. IEEE Communications Surveys & Tutorials, 2012, 15(1): 255-270.
4. H. Ghazzai, E. Yaacoub, M. Alouini, Z. Dawy and A. Abu-Dayya, "Optimized LTE Cell Planning With Varying Spatial and Temporal User Densities," in IEEE Transactions on Vehicular Technology, vol. 65, no. 3, pp. 1575-1589, March 2016, doi: 10.1109/TVT.2015.2411579.
5. F. Gordejuela-Sanchez and J. Zhang, “LTE access network planning and optimization: A service-oriented and technology-specific perspective,” in Proc. IEEE GLOBECOM, Honolulu, HI, USA, Dec. 2009, pp. 1–5.
6. A. Guo and M. Haenggi, “Spatial stochastic models and metrics for the structure of base stations in cellular networks,” IEEE Trans. WirelessCommun., vol. 12, no. 11, pp. 5800–5812, Nov. 2013.
7. 周志华.机器学习[M].清华大学出版社,2016.
8. 李航.统计学习方法[M].清华大学出版社,2012.