

doi:10.14132/j.cnki.1673-5439.2020.06.014

# 基于大数据挖掘的 LTE 网络重叠覆盖优化方法

张吉, 赵夙, 朱晓荣

(南京邮电大学 江苏省无线通信重点实验室, 江苏 南京 210003)

**摘要:**随着无线网络的快速发展,网络中重叠覆盖现象越来越严重,重叠覆盖区域频繁切换,导致系统容量减小,增加了掉话的可能,极大降低受影响区域的用户感知性能,因此重叠覆盖是网络结构优化中的研究重点。文中基于南京市江宁地区的实际路测数据,提出了基于大数据挖掘的 LTE 网络重叠覆盖优化方法。首先,对采集到的数据进行预处理和扩充,然后使用随机森林算法提取产生重叠覆盖的重要参数,基于区域重叠覆盖率对该参数进行调节,并使用支持向量机算法预测参数调节后的区域重叠覆盖率。仿真实验结果表明,该方法有效降低了区域的重叠覆盖率。

**关键词:** LTE; 重叠覆盖; 数据挖掘; 功率调节

**中图分类号:** TP391 **文献标志码:** A **文章编号:** 1673-5439(2020)06-0092-08

## Optimization method for overlapping coverage of LTE networks based on big data mining

ZHANG Ji, ZHAO Su, ZHU Xiaorong

(Jiangsu Key Laboratory of Wireless Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:** With the rapid development of wireless networks, the overlapping coverage of LTE networks is becoming more and more serious. Frequent handovers in overlapping coverage areas reduce the system capacity, increase the possibility of dropped calls and degrade the perceived performance of users in the areas affected by the overlapping coverage. Therefore, the overlapping coverage is the research focus in network structure optimization. According to the actual road test data in Jiangning New Area of Nanjing, an optimization method for the overlapping coverage in LTE networks based on big data mining is proposed. Firstly, the collected data are preprocessed and expanded. Then, the important parameters resulting in overlapping coverage are extracted by using a random forest algorithm. The parameters are adjusted based on the regional overlapping coverage rate. Finally, the support vector machine is used to predict the area overlapping coverage rate after adjustment. Simulation results show that the method can effectively reduce the overlapping coverage rate of the area.

**Keywords:** LTE; overlapping coverage; data mining; power adjustment

收稿日期: 2020-02-26; 修回日期: 2020-05-16 本刊网址: <http://nyzr.njupt.edu.cn>

基金项目: 江苏省无线通信重点实验室开放研究基金(2019WICOM01)、国家自然科学基金(61871237)和江苏省高校自然科学研究重大项目(16KJA510005)资助项目

作者简介: 张吉, 女, 硕士研究生; 赵夙(通信作者), 女, 副教授, zhaos@njupt.edu.cn

引用本文: 张吉, 赵夙, 朱晓荣. 基于大数据挖掘的 LTE 网络重叠覆盖优化方法[J]. 南京邮电大学学报(自然科学版), 2020, 40(6): 92-99.

为了增加频谱利用率,第四代移动通信网络采取同频组网的方式。随着基站间的间距逐渐缩短,同频小区之间出现重叠覆盖的现象越来越严重。重叠覆盖区域网络吞吐量下降,用户频繁切换掉话,不仅导致用户体验感降低,而且增加了运行商的网络建设和维护成本。因此,重叠覆盖优化一直是网络结构优化的重要组成部分<sup>[1]</sup>。

重叠覆盖主要源于不合理的网络结构,这里的网络结构包括基站站址、基站高度、基站间隔、天线方位角、参考信号发射功率等。重叠覆盖可能由其中的某一因素影响造成,也可能由多种因素共同影响造成。主要包括:(1)高站低下倾角,密集市区的站间距小于300 m,较小的站间距加上较低的下倾角,会在一个区域内出现较多的小区信号,产生重叠覆盖。(2)天线性能异常,天线老化或者故障,导致天线旁瓣、后瓣信号泄漏严重,信号泄漏区域造成重叠覆盖。(3)主服务小区参考信号较弱,导致终端主服不明显或者无主覆盖小区,产生重叠覆盖现象<sup>[2]</sup>。

传统网络优化方法,采用人工实地勘测调优,随着建网规模越来越大,这种方法效率低,难以满足现网基站优化的需求。

针对网络覆盖优化问题,许多文献提出了与传统方法不同的天线参数优化方案。文献[3]使用遗传算法对天线的方位角、下倾角和功率等参数同时进行调节,并通过数学模型计算调整后的采样点的覆盖情况。文献[4]认为可通过调节基站倾角的方法增强用户接收的有用信号的功率,

并在不同的无线信道模型下使用基于强化学习算法,实现了基站倾角的自动调节。文献[5]使用可变长粒子群算法调节天线的下倾角,在考虑容量约束的同时,满足覆盖条件。文献[6]针对弱覆盖和重叠覆盖问题,使用机器学习的思想联合调制下倾角和方位角。

上述文献主要是根据假设的场景,提出天线参数优化算法。虽然仿真结果符合预期效果,但是现实场景比仿真环境复杂很多。本文从真实网络环境中获取数据,将大数据分析方法运用到网络规划中,将模型驱动转换为数据驱动,解决重叠覆盖问题。

## 1 系统模型

所谓重叠覆盖,是针对某个采样点处的用户设备(User Equipment, UE)的,首先,UE接收到的主服务小区的信号强度 $RSRP_{cell}$ 和邻小区的信号强度 $RSRP_{ncell_i}$ ( $i = \{1, 2, \dots, h\}$ )均大于等于 $-105$  dBm,即 $RSRP_{cell} \geq -105$  dBm,且 $RSRP_{ncell_i} \geq -105$  dBm,表明主服务小区和邻小区都可以提供接入服务。其次,主服务小区与邻小区频点相同,且主服务小区和邻小区的信号强度相差在6 dB以内,即 $EARFCN_{cell} = EARFCN_{ncell_i}, |RSRP_{cell} - RSRP_{ncell_i}| \leq 6$  dB。如果满足上述条件的小区个数大于等于3(包括主服务小区在内),则可以判断,在该采样点处发生了重叠覆盖<sup>[7-8]</sup>。

针对LTE网络重叠覆盖场景,本文提出的基于大数据挖掘的重叠覆盖优化系统模型如图1所示。

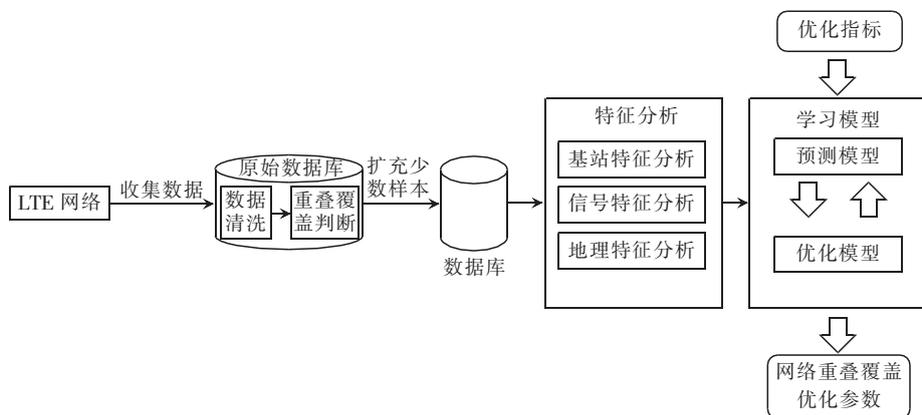


图1 基于大数据挖掘的重叠覆盖优化系统框图

首先,获得蜂窝网络实测数据,主要包括基站基本信息数据、终端测试点数据和地理位置数据。对数据进行清洗,去除缺省项和重复值,并根据重叠覆盖概念对数据进行重叠覆盖判断。因为重叠覆盖数

据在数据集中只占据了少数,所以这是一个典型的不平衡数据集,传统的学习方法以提升总体分类精度为目标,分类器在多数类的分类精度较高而在少数类的分类精度很低,为避免这一现象的产生,本文

使用 SMOTE 算法扩充数据集中的少数样本,使得数据集成为一个平衡数据集。对扩充后的平衡数据集运用随机森林算法进行特征分析。结合实际网络优化操作和特征重要性排序结果,本文认为优化基站功率会对降低区域的重叠覆盖率产生积极的影响。因此,在学习模型中根据各个小区的重叠覆盖率对天线的功率进行调整。如图 1 所示,学习模型由预测模型和优化模型两部分组成,每个小区的天线功率根据小区的重叠覆盖率进行调节。因为每次调节天线功率,重叠覆盖情况都会发生改变,所以,通过支持向量机算法学习原始数据,得到预测模型,对调整功率后的采样点信息进行重叠覆盖预测。优化模型则是根据预测得到的重叠覆盖情况,调整天线功率。整个模型的优化指标是,根据每次的重叠覆盖率逐步调节天线功率,直至重叠覆盖率不再降低。

## 2 数据格式和数据预处理

### 2.1 数据格式和清洗

城市中进行外卖服务和快递服务的工作人员,每天都会经过大量的道路,将路测仪器发放给这部分工作人员,在他们进行本职工作时,可以帮助运营商或者数据公司获得需要的路测数据。本文通过此方法获得南京江宁地区 2019 年 7 月 24 日至 2019 年 7 月 30 日部分时间段的蜂窝网络数据,总计 111 941 条。基站侧数据从网管中提取。基站侧和采样点侧属性如表 1 和表 2 所示。

表 1 基站侧属性

属性名	说明
E-CGI	全球小区识别号
TAC	跟踪区
PCI	物理扇区标识
LNG	基站经度
LAT	基站纬度
Bore	天线方位角
Tilt	天线下倾角
tilt-M	机械下倾角
tilt-E	电子下倾角
High	基站站高
Power	天线发射功率
Up_throughput	上行吞吐量
Down_throughput	下行吞吐量

表 2 采样点侧属性

属性名	说明
Collect-Time	采集时间
IMEI	设备号
LAT	采样点纬度
LNG	采样点经度
ECI	小区编号
EARFCN	主服务小区频点
PCI	主服务小区物理小区标识
RSRP	主服务小区参考信号强度
EARFCN_1	邻区 1 频点
PCI_1	邻区 1 物理小区标识
RSRP_1	邻区 1 参考信号强度
EARFCN_2	邻区 2 频点
PCI_2	邻区 2 物理小区标识
RSRP_2	邻区 2 参考信号强度

在基站侧属性中,TAC 用于手机定位,PCI 用于 LTE 终端区分不同小区的无线信号,这两个属性与重叠覆盖无关,所以剔除。E-CGI 由四部分组成:移动国家码、移动网络码、位置区号码和小区标识码。在采样点侧属性中,ECI 为小区编号,提取基站侧属性 E-CGI 中的位置区号码和小区标识码字段,通过 ECI 将采样点数据与对应的基站匹配。部分采样点处信号良好,主服务小区信号完全覆盖且没有受到邻小区信号的干扰,这些采样点没有测量得到邻区 1 和邻区 2 的信息,也不会产生重叠覆盖现象。

在实际采样活动中,外卖人员和快递服务人员经常会在一段连续的时间内在同一地点停留,导致获得某个地点大量的重复数据,在初步清洗时,删除这部分重复数据。对保留的采样点数据根据重叠覆盖定义进行重叠覆盖判断,发生重叠覆盖的采样点打上标签 1,没有发生重叠覆盖的采样点,打上标签 0。

通过基站的位置信息和采样点的位置信息,计算出两者之间的距离,作为新特征 DIFF 记录到表格。最终选择的特征包括:基站位置信息(包括经纬度)、采样点位置信息(包括经纬度)、基站和采样点距离信息、基站方位角、基站下倾角、电子下倾角、物理下倾角,基站站高、基站上下行吞吐量。

在江宁地区北纬  $31.770^{\circ} \sim 31.784^{\circ}$ ,东经  $118.82^{\circ} \sim 118.862^{\circ}$  范围的地理区域内(面积大约是  $4.2256 \text{ km}^2$ ),发生了较严重的重叠覆盖现象,该区域的基站位置和重叠覆盖采样点的地理位置绘制如

图2所示。在图2中,红色星形表示重叠覆盖采样点,绿色星形表示基站。

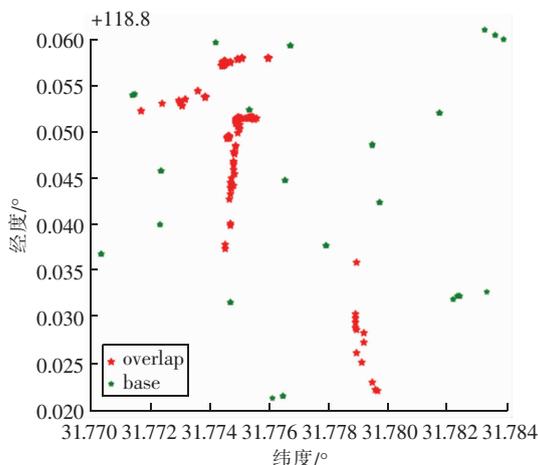


图2 初始重叠覆盖区域地理图像

图2中包含了三个重叠覆盖严重区域,用不规则四边形逼近这些区域,计算出重叠覆盖面积。区域一大约是 $0.089 \text{ km}^2$ ,区域二大约 $0.14 \text{ km}^2$ ,区域三大约是 $0.023 \text{ km}^2$ 。

图2地理区域内建设了19个宏基站,但是实际只对采样点附近的8个宏基站进行了信息采集,这8个宏基站的基站编号,地理位置信息如表3所示。

表3 宏基站信息

序号	eNodeB-ID	LNG_B/°	LAT_B/°
1	38 184	118.848 640	31.779 444
2	855 398	118.845 810	31.772 380
3	36 514	118.852 401	31.775 316
4	36 752	118.821 527	31.776 441
5	629 784	118.831 595	31.772 380
6	855 549	118.844 830	31.776 540
7	860 315	118.832 250	31.782 380
8	900 605	118.837 740	31.777 890

## 2.2 数据集不平衡和扩充

经过数据匹配、清洗,最终得到6 881条数据,其中标签为0的采样点数据量为939条,占比13.6%,标签为1的采样点数据量为5 941条,占比86.3%。这是一个典型的不平衡数据集。以二元分类为例,数据集不平衡意味着数据集中一类的样本个数多于另一类,即数据集中存在着多数类样本和少数类样本<sup>[9]</sup>。分类学习算法以总体分类精度最大为目标会使得分类模型偏向于多数类样本,对于那些旨在检测罕见但重要的案例的实际应用领域,例如欺诈检测、贷款违约检测和癌症检测,不平衡数

据集会使算法的性能大大下降<sup>[10]</sup>。扩充少数类样本的个数,即过采样技术,一直是解决不平衡数据集的一个有效方法。

人工合成少数类过采样技术(Synthetic Minority Over-Sampling Technique, SMOTE)是解决数据集不平衡问题的最流行和最著名的采样算法之一,与随机过采样技术相比,它可以在很大程度上减轻过拟合的问题<sup>[11]</sup>。SMOTE算法的基本思想是利用少数类样本的某个样本及其近邻样本合成新的样本。用 $D_{\min}$ 表示少数类样本集合, $(x_i, y_i)$ 表示 $D_{\min}$ 中的某一个样本,其 $K$ 近邻同类样本表示为: $X' = \{(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_n, y'_n)\}$ ,算法具体流程可分为三步:

(1) 对于 $D_{\min}$ 中的每一个样本 $x_i$ ,以欧氏距离为标准计算它到 $D_{\min}$ 中每一个样本的距离,得到其 $K$ 近邻。

(2) 确定采样倍率 $N$ ,根据采样倍率在 $x_i$ 的 $K$ 近邻中选择几个样本,假设选择的样本为 $x'_j$ 。

(3) 在 $x_i$ 和 $x'_j$ 之间进行线性插值合成新样本,合成新样本的公式如下:

$$x_{\text{new}} = x_i + \text{rand}(0, 1) \times (x'_j - x_i)$$

## 3 最优基站功率部署

### 3.1 特征重要性和特征选择

在数据挖掘中,特征选择在处理高维数据方面一直发挥着重要作用。特征选择技术从数据集的原始特征中选择特征子集,减少特征间的冗余,提高机器学习中的分类问题的准确性<sup>[12]</sup>。近年来,基于随机森林算法的特征选择已经在很多领域发挥了重要作用,该方法的优点如下:(1) 基于随机森林的特征选择在训练过程中自动执行;(2) 由于在构建过程中应用了随机选择,因此特征选择结果具有很强的归纳能力和较高的准确性;(3) 基于随机森林的特征选择可以为特征提供排名结果<sup>[13]</sup>。因此,本文通过随机森林算法给出特征重要性分数,在了解产生重叠覆盖原因的同时,将特征重要性结果和实际网络调优工作的复杂度相结合,选择最合适的天线参数进行调节。

随机森林算法是集成学习算法的一种。随机森林算法的随机性体现在两个方面。第一,在以CART决策树为基学习器构建Bagging集成的基础上,进一步在决策树的训练过程引入了随机属性选择。第二,从原始数据集中采取放回的抽样方式构造子数据集。CART决策树使用基尼系数选择划分

属性。因此,随机森林算法在构造基决策树的同时,根据特征节点划分前后基尼系数的改变,给出变量重要性评分(Variable Importance Measures, VIM)。

使用随机森林获得特征重要性的具体算法流程如下:

输入:训练集  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$

样本特征数  $M$

决策树算法  $\zeta$

训练轮数  $T$

过程:

1. for  $t = 1, 2, \dots, T$  do

2. 从训练集  $D$  中以拿取又放回的形式随机采样  $N$  个样本,构成一个新的训练集

3. 从总量为  $M$  的特征向量中随机选择  $m$  个特征

4.  $h_t = \zeta(N, m)$

5. End for

6.  $H(x) = \arg \max_{y \in Y} \sum_{t=1}^T I(h_t(x) = y)$

输出:特征的重要性排序

表 4 中记录了特征重要性排名前七的特征名称、说明及其重要性评分。

表 4 前七个特征的重要性

特征	说明	重要性
DIFF	基站和采样点的距离	0.218
LAT	采样点纬度	0.217
LNG	采样点经度	0.154
Bore	天线方位角	0.128
Power	天线功率	0.113
th_down	下行吞吐量	0.071
th_up	上行吞吐量	0.067

在表 4 中,特征重要性排名前三的特征均与基站和采样点的地理位置信息有关,即基站和采样点的距离是造成重叠覆盖的主要原因。其次是天线方位角和天线功率因素,且天线方位角和天线功率的重要性评分只相差了 0.015。

在现网调优中,移动原始基站位置需要重新勘测周围的环境,耗费巨大的人力物力,且调整基站位置属于网络规划而不是网络优化,所以不对基站位置调优进行考虑。调整天线方位角需要人工上山,实施起来比较困难。而对基站功率进行调节,只需要从后台操作,实施容易。采样点发生重叠覆盖主要是因为接收到的主服务小区的参考信号强度接近于邻小区,导致切换频繁,用户体验度低。采取增强主服务小区的天线功率的方法,可使得重叠覆盖处恢复正常覆盖。

### 3.2 分类性能评估

针对二分类问题,通常使用正确率(Accuracy, ACC),精准率(Precision, PRE),召回率(Recall, Rec),F1-score 等评价指标评估模型的性能。这些评价指标建立在混淆矩阵上,表 5 记录了混淆矩阵内容。

表 5 混淆矩阵

分类	实际正类	实际负类
预测正类	TP	FP
预测负类	FN	TN

表 5 中,TP 表示实际正类预测为正类的个数,TN 表示实际负类预测为负类的个数,FP 表示实际负类预测为正类的个数,FN 表示实际正类预测为负类的个数。

根据混淆矩阵,给出正确率,召回率,F-score 的定义式:

正确率:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

正确率表示预测正确的样本个数与所有拿来预测的样本个数的比值。在一些样本极度不均衡的情况下,正确率没有任何意义。例如,在某个二分类样本集中,样本的总个数为 100,其中正类个数为 95,负类个数为 5。如果算法将所有的样本全部预测为正类,那么算法的预测正确率高达 0.95。但实际该算法不具有辨别负类的能力。此时,需要更换对算法预测结果的评价标准。

精准率:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

精准率表示预测为正类的样本中,实际为正类样本的占比。例如,precision = 0.7 表示分类器预测为正例的结果中,只有 70% 是真正的比例,即对于一个预测为正例的结果,只有 70% 的可能性是正确的。

召回率:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

召回率表示实际正类样本中预测为正类的样本的占比。Recall = 0.7,表示分类器只能查出 70% 的正例。

根据不同的研究场景,使用精准率和召回率两种评价标准中的一种。例如在评价算法预测假币的性能中,更看重精准率;在评价算法预测病患的性能

中,更看重召回率,因为在这种情况下,医生更看重对实际存在的病例没有做出错误的预测。

精准率和召回率一般不是正相关的关系,使用F1-score评价指标兼顾精准率和召回率。定义如下:

$$\frac{1}{F_1} = \frac{1}{2} \left( \frac{1}{\text{precision}} + \frac{1}{\text{recall}} \right) \quad (4)$$

F1-score是精准率和召回率的调和平均数,只有当精准率和召回率都非常高时,F1-score的值才会高,如果其中一个评价指标的值低,F1-score的结果就会接近值比较低的评价指标。

### 3.3 重叠覆盖预测模型

本文使用支持向量机(Support Vector Machine, SVM)算法作为重叠覆盖预测模型。SVM算法基于统计学习理论,通过构造两个平行的超平面,以使两个类别分开,并使两个超平面之间的距离最大,因此,SVM算法具有良好的泛化性能,在分类和回归问题上具有出色的表现,被广泛应用于各类实际问题中<sup>[14]</sup>。对于标准线性支持向量分类算法(Support Vector Classification, SVC),分离两种样本的超平面可以定义为:

$$f(x) = w^T x + b = 0 \quad (5)$$

通过引入正则项和松弛变量 $\zeta$ ,可以将优化问题描述如下:

$$\arg \min_{w, b, \zeta} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \zeta_i \quad (6)$$

$$\text{s. t. } y_i(w^T x_i + b) \geq 1 - \zeta_i, \zeta_i \geq 0, \forall x_i$$

在实际操作中,并没有对这一优化问题进行求解,而是求解其对偶问题,获得决策函数。SVM算法通过核函数将数据映射到高维特征空间,解决样本线性不可分的问题。

### 3.4 功率优化算法

通过3.1节特征重要性分析得到,基站发射功率是影响基站覆盖情况的重要因素。重叠覆盖发生的主要原因是用户设备接收到的主服务小区的信号电平接近于邻小区的信号电平,导致用户设备发生频繁切换的情况。如果增强主服务小区的发射功率,可以使重叠覆盖处恢复正常覆盖,提升用户体验。

本文以小区重叠覆盖率为指标调节小区功率,优化目标是使整体重叠覆盖率降低。小区重叠覆盖率和整体重叠覆盖率定义分别为

$$\text{ration}_i = \frac{\text{area}_{\text{overlapping}_i}}{\text{area}_i} \quad (7)$$

$$\text{Ration} = \frac{\sum_{i=1}^N \text{area}_{\text{overlapping}_i}}{\text{Area}} \quad (8)$$

在式(7)中, $\text{ration}_i$ ,  $\text{area}_{\text{overlapping}_i}$ ,  $\text{area}_i$ 分别表示第*i*个小区的重叠覆盖率,重叠覆盖面积和总面积。在式(8)中,Ration, Area表示整体重叠覆盖率和总面积,*N*表示小区总个数。

本文通过对采样点划分栅格的方式计算小区面积和重叠覆盖面积。路测数据中采样点的经纬度信息可精确到小数点后第七位,经纬度小数点后第四位可精确到13 m。对比每一个采样点的经度和纬度,将经度和纬度前四位相同的采样点划入同一个栅格,因此一个栅格的面积为13 m × 13 m。当栅格中重叠覆盖采样点数在总采样点数中占比超过50%时,将该栅格判定为重叠覆盖栅格。

本文以0.2 dBm为步长对每个小区的发射功率进行调节,即每个小区调节的功率幅度为:

$$\text{power}'_i = \text{power}_i + \text{ration}_i \times 0.2 \quad (9)$$

其中, $\text{power}_i$ ,  $\text{power}'_i$ 分别表示小区*i*优化前后的天线功率值。

将进行功率调节后的采样点的特征输入重叠覆盖预测模型进行预测,根据预测结果计算当前全局的重叠覆盖率。当全局的重叠覆盖率不再降低时,停止迭代。整个规划过程的总目标函数如式(10)所示:

$$\min \text{Ration} = \frac{\sum_{i=1}^N \text{area}_{\text{overlapping}_i}}{\text{Area}} \quad (10)$$

算法过程如下:

输入:训练样本集

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

$$x_i \in X \subseteq \mathbf{R}^n, y_i \in Y = \{0, 1\}$$

过程:

1. 计算初始重叠覆盖率  $\text{overlapping}_{\text{initial}}$
2. 将训练样本输入重叠覆盖预测模型进行模型训练,返回训练好的模型  $y_i = \text{SVM}(x_i)$
3. 根据采样点的小区编号,将编号相同的采样点放入同一个小区,根据本小区的重叠覆盖率计算增加功率的值,并对小区功率进行调节。
4. 将产生的新样本输入预测模型,进行重叠覆盖判断。
5. 计算全局的整体覆盖率,如果此时的重叠覆盖率不再降低,则认为达到优化目标,停止计算,否则,继续重复步骤3至5。

输出:达到降低重叠覆盖目标时,各个小区的功率值。

## 4 仿真结果

本实验数据来源于8个宏基站和经过数据清

洗、扩充后的 11 745 条测试点的路测数据,包含基站相关信息数据、终端采样点数据以及基站和采样点的地理位置数据,仿真结果验证了本文提出的优化重叠覆盖方法的可行性,并且采用 Python Matplotlib 工具将实验结果可视化。

#### 4.1 数据集扩充结果与分析

使用 SMOTE 算法对原始数据集的少类样本进行扩充,SMOTE 算法默认扩充后,两类样本的个数相同。但是观察生成的数据集,部分生成数据不符合要求。例如,某一个采样点的方位角数值为 280.897 778,而真实的方位角的数值为 280,将这部分生成数据定义为噪声数据,并且剔除。

表 6 记录了经过 SMOTE 算法扩充以后两类样本集包含的样本个数以及两类样本占比。图 3 绘制了扩充样本以后的重叠覆盖区域地理图像。与图 2 初始重叠覆盖地理图像对比,扩充的重叠覆盖采样点位于原始三个重叠覆盖区域处,证明 SMOTE 算法扩充的数据的合理性。

表 6 SMOTE 算法扩充后的两类样本个数

样本类型	样本数量	样本占比
标签为 0	5 825	0.495
标签为 1	5 920	0.504

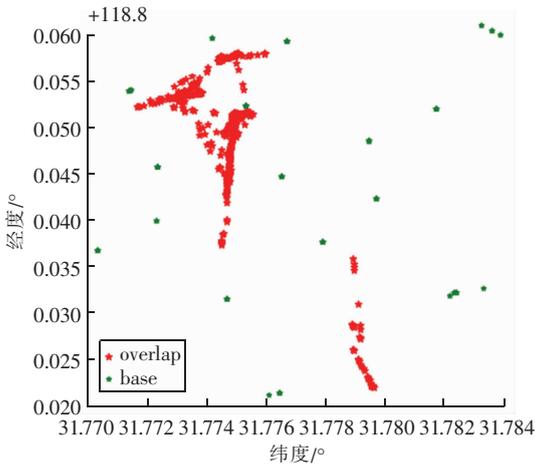


图 3 扩充样本以后重叠覆盖区域地理图像

#### 4.2 功率调整结果和分析

图 4 为功率优化后的重叠覆盖地理图像,与图 3 相比,三个重叠覆盖区域均有了改善。图 3 中的重叠覆盖率为 45%,图 4 中的重叠覆盖率为 27%,说明通过优化算法,该区域内的重叠覆盖情况改善幅度达到 18%。

表 7 记录了每个小区优化前后的天线功率参数和功率调整幅度。表 8 记录了每个小区优化前后的重叠覆盖率。

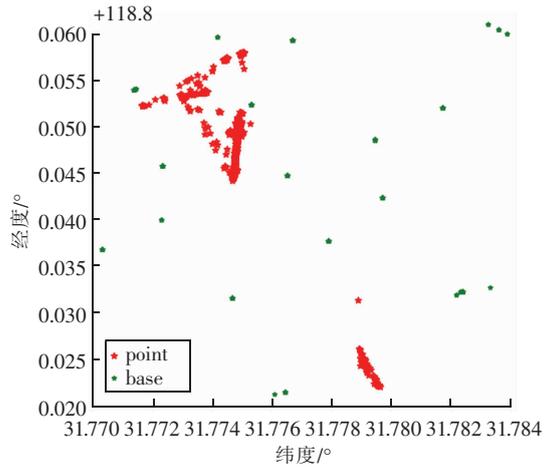


图 4 功率优化后的重叠覆盖地理图像

表 7 各个小区优化前后天线功率参数

序号	ECI	初始功率	优化后的功率	功率调整幅度
		/0.1dBm	/0.1 dBm	/0.1 dBm
1	855 491	122	131	9
2	855 494	92	109	17
3	855 495	92	109	17
4	855 496	92	109	17
5	860 315 3	122	135	13
6	860 315 5	122	122	0
7	900 605 5	92	108	16
8	900 605 6	122	122	0
9	365 141 51	122	140	18
10	365 141 53	132	149	17
11	365 141 54	122	122	0
12	365 141 55	122	123	1
13	365 141 58	122	122	0
14	365 141 60	122	122	0
15	367 521 51	152	160	8
16	860 315 16	122	122	0
17	627 984 154	139	139	0

表 8 各个小区优化前后重叠覆盖率

序号	ECI	初始重叠覆盖率	优化后的重叠覆盖率
1	855 491	0.714	0.571
2	855 494	0.667	0.667
3	855 495	0.672	0
4	855 496	0.667	1.000
5	860 315 3	1.000	0
6	860 315 5	0	0
7	900 605 5	1.000	0.778
8	900 605 6	0	0
9	365 141 51	1.000	1.000
10	365 141 53	0.500	0.937
11	365 141 54	0.375	0.125
12	365 141 55	0.536	0.214
13	365 141 58	0.222	0.111
14	365 141 60	0.286	0.429
15	367 521 51	0.429	0.857
16	860 315 16	0.061	0
17	627 984 154	0.143	0

在表7中,功率调整幅度为0~1.8 dBm,调整范围较小,在没有过多增加基站功耗的同时有效降低了区域的重叠覆盖情况。在表8中,17个小区中有12个小区的重叠覆盖率得到了降低,而855 494、855 496、365 141 51、365 141 53、367 521 51这5个小区的重叠覆盖率没有得到降低甚至有了升高,这一现象符合实际情况,即重叠覆盖现象可能由多个因素造成,需要联合调节基站方位角和下倾角,后续对联合调节问题继续进行讨论。

## 5 结束语

针对无线网络重叠覆盖场景,本文提出了基于大数据挖掘的重叠覆盖优化方法。首先利用路测数据对整体网络的重叠覆盖情况进行初步分析,利用随机森林算法选取影响重叠覆盖的重要特征。然后,基于区域重叠覆盖率调整天线参数,并通过支持向量机算法构建预测模型,对调整参数后的数据进行重叠覆盖判断。

实验结果表明,本文提出的基于大数据挖掘的网络覆盖优化方法,是一种切实可行的LTE网络覆盖优化方法。该方法依托于实际网络场景,与针对仿真场景提出的天线参数优化方法相比,更具有现实指导意义。但是在实验结果中,部分小区的重叠覆盖率没有下降甚至有了升高,说明对这部分小区仅仅调节功率一个参数,结果并不理想,需要联合调节其他参数,这是下一步研究的方向。

### 参考文献:

- [1] 赵明峰,黄建辉,梁金山,等. 基于MR与扫频数据的LTE-FDD重叠覆盖优化方法[J]. 电信科学,2019,35(S1):124-128.  
ZHAO Mingfeng, HUANG Jianhui, LIANG Jinshan, et al. LTE-FDD overlapping coverage optimization method based on MR and scanning data[J]. Telecommunications Science, 2019, 35(S1):124-128. (in Chinese)
- [2] 武海斌. TD-LTE网络重叠覆盖分析[J]. 计算机与网络,2013,39(18):64-67.  
WU Haibin. Analysis on TD-LTE network overlapping coverage[J]. Computer & Network, 2013, 39(18):64-67. (in Chinese)
- [3] 谷欣杏. LTE网络覆盖优化及无线定位优化算法的研究[D]. 北京:北京邮电大学,2019.

- GU Xinxing. Research on LTE network coverage optimization and wireless positioning algorithm[D]. Beijing: Beijing University of Posts and Telecommunications, 2019. (in Chinese)
- [4] DANDANOV N, SAMAL S R, BANDOPADHAYA S, et al. Comparison of wireless channels for antenna tilt based coverage and capacity optimization[C]//The 6th Global Wireless Summit. 2018.
- [5] PHAN N, BUI T, JIANG H L, et al. Coverage optimization of LTE networks based on antenna tilt adjusting considering network load[J]. China Communications, 2017, 14(5):48-58.
- [6] LIN Zhengyi, OUYANG Ye, SU Le, et al. A machine learning assisted method of coverage and capacity optimization (CCO) in 4G LTE self organizing networks (SON)[C]//Wireless Telecommunications Symposium. 2019:1-9.
- [7] 崔航,王四海,李新,等. TD-LTE重叠覆盖及解决方案分析[J]. 移动通信,2013,37(21):17-21.
- [8] 杜杨,杨奥林,朱革. TD-LTE网络中小小区重叠覆盖优化方法研究[J]. 移动通信,2017,41(1):74-77.  
DU Yang, YANG Aolin, ZHU Ge. Investigation on the cell overlapping coverage optimization in TD-LTE networks[J]. Mobile Communications, 2017, 41(1):74-77. (in Chinese)
- [9] LIU Y, WANG Y Z, REN X G, et al. A classification method based on feature selection for imbalanced data[J]. IEEE Access, 2019, 7:81794-81807.
- [10] FENG L, WANG H B, JIN B, et al. Learning a distance metric by balancing KL-divergence for imbalanced datasets[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 49(12):2384-2395.
- [11] CHENG K, ZHANG C, YU H L, et al. Grouped SMOTE with noise filtering mechanism for classifying imbalanced data[J]. IEEE Access, 2019, 7:170668-170681.
- [12] HARIHARAN S, MANDAL D, TIRODKAR S, et al. A novel phenology based feature subset selection technique using random forest for multitemporal PolSAR crop classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 11(11):4244-4258.
- [13] PENG X S, LI J S, WANG G J, et al. Random forest based optimal feature selection for partial discharge pattern recognition in HV cables[J]. IEEE Transactions on Power Delivery, 2019, 34(4):1715-1724.
- [14] VAPNIK V. The Nature of Statistical Learning Theory[M]. New York:Springer, 1995.