基于LightGBM算法的MR网络信号预测

MR Network Signal Prediction Based on LightGBM Algorithm

张硕伟[1]，裴明丽[2]，高有利[1]，黄　铭[1]，刘贤松[1]（1. 中国联通网络AI中心，上海 200050；2. 科大国创软件股份有限公司，安徽合肥 230000）

Zhang Shuowei[1]，Pei Mingli[2]，Gao Youli[1]，Huang Ming[1]，Liu Xiansong[1]（1. China Unicom Network AI Center，Shanghai 200050，China；2. USTC Sinovate Software Co.，Ltd.，Hefei 230000，China）

摘　要：

无线网络质量是运营商的命脉与核心，合理的无线网络规划能够保障网络覆盖范围广、干扰低、系统服务质量高、运营成本低以及网络结构达到最佳。提出了一种基于LightGBM算法的网络信号预测方法，采集基站现有的MR数据，对MR数据进行栅格化，并对栅格化后的数据进行分析处理，通过LightGBM算法对有经纬度的未知地区进行网络信号强度预测，从而实现无线网络优化，保证网络通信质量。

关键词：

LightGBM算法；信号预测；MR栅格化；网络优化

doi：10.12045/j.issn.1007-3043.2020.10.005

文章编号：1007-3043（2020）10-0021-05

中图分类号：TN915

文献标识码：A

开放科学（资源服务）标识码（OSID）：

Abstract：

Wireless network quality is the core of telecom operators. Reasonable wireless network planning can guarantee wide network coverage，low interference，high system service quality，low operation cost and optimal network structure. It proposes a network signal prediction method based on lightGBM algorithm，which collects the existing MR data of the base station and rasterizes the MR data，then the rasterized data is analyzed and processed，and the network signal strength of the unknown area with longitude and latitude is predicted by LightGBM algorithm，so as to optimize the wireless network and ensure the quality of network communication.

Keywords：

LightGBM algorithm；Signal prediction；MR rasterization；Network optimization

引用格式：张硕伟，裴明丽，高有利，等. 基于LightGBM算法的MR网络信号预测[J]. 邮电设计技术，2020（10）：21-25.

1 概述

随着LTE网络大规模应用以及市场竞争的白热化，用户越来越重视自身的感知体验，因此运营商对覆盖优化和质量优化的要求也越来越高[1]。现阶段的无线网络优化工作[2]，主要采用CQT（Call Quality Test）、DT（Driving Test）和用户投诉等方式发现覆盖问题和质量问题[3-6]。但是CQT和DT方式需要运营商投入大量的时间和人力，用户投诉方式又严重影响用户感知和满意度。

收稿日期：2020-08-26

针对传统方法存在的弊端，本文提出了一种基于LightGBM算法的网络信号预测的新方法，使用MR数据和LightGBM算法对未知地区的网络信号进行预测，不仅解决了现有技术数据采集成本高、数据分析过程烦琐等问题，还创新地将AI技术与网络优化相结合，提高无线网络优化的自动化水平。

2 LightGBM算法介绍

在机器学习算法领域，监督学习算法中最常用的2类算法为回归（Regression）算法和分类（Classification）算法[7]。回归算法和分类算法的区别在于输出变量的类型不同，定量输出或者连续变量预测称为“回

归”；定性输出或者离散变量预测称为“分类”[8]。而对网络信号预测过程是一个典型的回归问题，因此，可以利用回归算法对网络信号进行精准预测。目前比较流行的回归算法是集成学习Boosting算法中的梯度提升树（GBDT——Gradient Boosting Decision Tree）算法[9-10]和极端梯度提升（XGBOOST——eXtreme Gradient Boosting）算法[11-12]。其中，GBDT算法是一种迭代的决策树算法，该算法由多棵决策树组成，所有决策树的结果累加起来做为最终结果。在机器学习领域中，GBDT是一个经久不衰的模型：

GBDT = Gradient Boosting + Decision Tree

GBDT具有Gradient Boosting和Decision Tree的功能特性，主要优点是训练效果好、不易过拟合且泛化能力较强。通过多轮迭代，每轮迭代产生一个弱分类器，后续每个分类器在上一轮分类器的残差基础上进行训练，如图1所示。

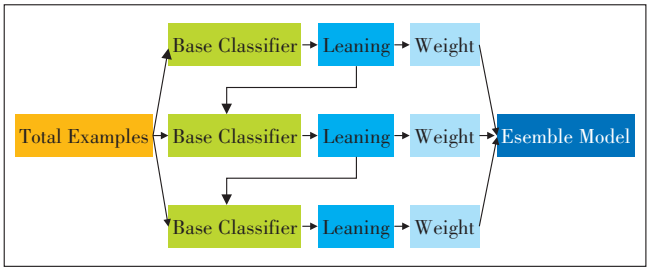

图1　GBDT的模型训练过程

XGBOOST算法是GBDT的改进，它是大规模并行boostedtree的工具，是目前最快最好的开源boostedtree工具包。在XGBOOST之后，微软公司又提出了一种LightGBM算法来增强GBDT的性能[13]。LightGBM是一个实现GBDT算法的框架，主要用于解决GBDT在大规模数据处理上遇到的问题。采用带深度限制的Leaf-wise的叶子生长策略[14]，其计算代价小，且避免了过拟合。为了减小存储成本和计算成本，LightGBM算法是一种基于Histogram的决策树算法。此外LightGBM直接支持类别特征处理，使其性能得到较好的提升。因此，基于以上集成机器学习算法优劣势比较，提出了一种基于LightGBM算法的网络信号预测方法。

3　一种网络信号的预测方法

为了提高算法模型预测的准确性，本文采用了LightGBM机器学习算法和迭代优化的训练方式。该模型算法的整体流程框架如图2所示，整个模型训练过程可分为5个步骤：数据收集、数据处理、模型训练、模型验证和精度评价。

3.1　数据采集与处理

本文采集了某市不同基站的MR原始样本数据，MR数据是一种测量报告，由用户终端周期性上报给基站控制器（包含小区下行信号强度、信号质量等信息），再由基站控制器收集和统计[15]。将采集到的MR数据映射到栅格上，得到栅格基本信息，包括位置信息和小区配置数据，其中位置信息包括区域、经度、纬度、位置类型（室内/室外）；小区配置数据包括基站位置、基站高度、小区方位角、工作频段、总下倾角、中心载频的信道号等。具体字段描述如表1所示。

本文提出一种子栅格的概念，将50×50栅格根据道路和楼栋的GIS边界进一步细分成子栅格。首先，将带有经纬度的MR数据进行异常数据清洗和室内室外用户识别；其次，将带有室内室外标签的MR数据映射到对应子栅格之中。同时，为便于对MR数据主邻小区计算处理，根据当前主邻小区的记录数，将单条MR记录拆分多条记录，新增主邻小区标识，包括中心

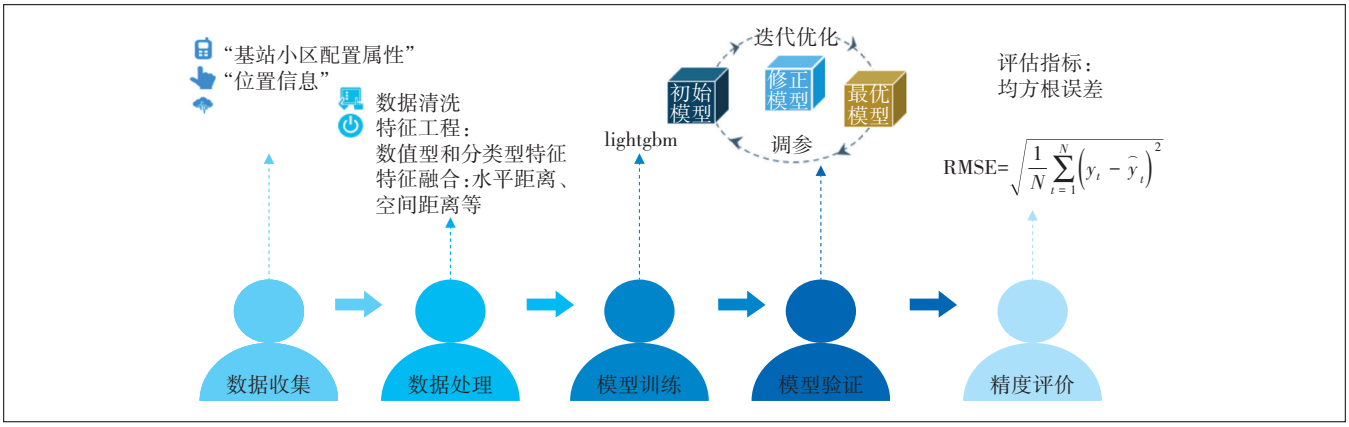

图2　模型整体流程图

表1　MR原始数据信息表

列名	描述	列名	描述
city_code	城市编码	cover_type	基站覆盖类型
lon	所在位置经度	jail_scene	小区场景
lat	所在位置纬度	work_frqband	工作频段
location_type	位置类型，室内/室外	azimuth_angle	方位角
cgi	基站唯一标识	total_downtilt	总下倾角
longitude	基站纬度	site_height	站高
latitude	基站纬度	Center_freq_point	中心载频的信道号
cover_scene	基站覆盖场景		

载频的信道号、物理小区识别码和识别邻区CGI等信息。最后，基于小区间和电平值间的相似性，利用K-means聚类算法将MR样本点分组，提升子区域内电平值的特征性，得到子栅格的基本信息。

子栅格信息具体字段信息如表2所示，其中字段rsrp能够用于判断是否需要调整小区的天线天馈角以及确定小区各位置的信号强度。

表2　子栅格信息表

列名	描述	列名	描述
city_code	城市编码	grid_sub_type	子栅格类型，1-室内 2-道路 3-室外
ta	时间提前量	group_id	分组
lon	子栅格的经度	If_main_cell	是否主小区
lat	子栅格的纬度	rsrp	电信号值
grid_id	栅格id	cgi	基站唯一标识
grid_sub_id	子栅格id	date	日期

栅格位置信息表反应了每个栅格中子栅格的具体位置信息，具体字段如表3所示。

表3　栅格位置信息表

列名	描述	列名	描述
city_code	城市编码	grid_b_r_x	右下的经度
grid_id	栅格id	grid_b_r_y	右下的纬度
grid_sub_id	子栅格id	grid_t_r_x	右上的经度
grid_sub_type	子栅格类型	grid_t_r_y	右上的纬度
group_xid	子栅格中心点坐标经度	grid_t_l_x	左上的经度
group_yid	子栅格中心点坐标纬度	grid_t_l_y	左上的纬度
grid_b_l_x	左下的经度	grid_sub_order_point	子栅格位置的顺序点
grid_b_l_y	左下的纬度		

3.2 特征工程

特征工程是机器学习研究课题中最重要的部分。在这一过程中需要找到最能反映分类本质的特征来完成原始数据的分类工作。总之，特征工程的研究是否精细，会直接影响到模型的预测性能。因此，需要对训练数据集进行特征处理。

3.2.1 处理无效值和缺失值

对MR原始样本数据集进行删除、去重和缺失值填充等处理，删除数据集中缺失经纬度以及group_id、date无关字段，去除数据集中重复数据以及中心载频的信道号(center_freq_point)中的空值以平均数填充。

3.2.2 文本数值化

MR原始样本数据集中有很多字段是文本形式，如位置类型(location_type)、基站覆盖场景(cover_scene)、基站工作频段(work frqband)、基站覆盖类型(jail_scene)、基站覆盖类型(cover_type)等字段。文本形式计算机无法识别，需要做数值化映射操作。

3.2.3 特征构造

首先，将子栅格经度lon、子栅格纬度lat、基站经度longitude、基站纬度latitude转化为对应的弧度值lon_1、lat_1、lon_2、lat_2，然后分别计算经纬度差值：

$$d_{lon}=lon_2-lon_1 \tag{1}$$

$$d_{lat}=lat_2-lat_1 \tag{2}$$

其次，计算空间距离，构造相应特征：

$$H_d = 2\times 6\,371\times 1\,000\times \sin^{-1}\sqrt{\sin\left(\frac{d_{lat}}{2}\right)^2+\cos(lat_1)\times\cos(lat_2)\times\sin\left(\frac{d_{lon}}{2}\right)^2} \tag{3}$$

最后，计算空间距离，构建相应特征：

$$S_d=\sqrt{(H_d)^2+(site\_height)^2} \tag{4}$$

3.3 模型训练

本文研究了包括GBDT、XGBOOST和LightGBM 3种最常用的机器学习算法的区别和特点，通过比较预测精度和复杂度，最终选择了LightGBM作为整个模型的核心算法，并通过Python编程语言实现数据处理和模型训练。其中，Python中的LightGBM参数设置如表4所示。

表4　模型参数设置表

参数名称	参数说明	参数取值
boosting_type	Boosting类型	GBDT
boosting	算法	dart
objective	模型用途	regression
metric	评估函数	RMSE
learning_rate	学习率	0.1
num_iteration	迭代次数	50 000

3.4 模型验证

为了保障模型的泛化性及精准度，将MR数据集划分为训练数据集和测试数据集，通过MR测试数据集验证模型对新的MR样本数据的判别能力，以测试误差作为模型泛化误差的近似值，最后选择泛化能力强的模型作为最终模型。本文采用留出法划分样本数据集，具体过程如下。

a）将MR样本数据集D划分为训练数据集X和测试数据集C，比例为9∶1。$X \cap C = \varnothing, X \cup C = D$。

b）将训练数据集X再次划分为模型训练数据集T和模型验证集Y，比例为8∶2。$T \cap Y = \varnothing, T \cup Y = X$。

c）为了保证训练和测试数据集的随机性，采用对MR样本数据集D多次划分的方式，每次数据集划分模型都会重新训练，计算每次模型训练的rsrp误差率，来反应模型预测效率。

$$\mathrm{rsrp}_{error} = \frac{\left|\mathrm{rsrp}_{预测值} - \mathrm{rsrp}_{真实值}\right|}{\mathrm{rsrp}_{真实值}} \times 100\% \qquad (5)$$

3.5 精度评估

为了保证模型训练后的精确度，采用均方根误差（RMSE——Root Mean Squared Error）来评估其预测精度[16]。RMSE的值越小，说明模型的预测结果越准确，即具有更好的精确度，RMSE公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t - \hat{y}_t\right)^2} \qquad (6)$$

式中：

N——观测次数

y_t——信号强度（rsrp）的真实值

$\hat{y}_t$——信号强度（rsrp）的预测值

4 实验分析

4.1 不同学习率对RMSE结果的影响

如图3所示，横坐标表示迭代次数，纵坐标表示RMSE值。在同样的迭代次数下，学习率为0.3时，RMSE的值最小，即模型预测效果越好。随着迭代次数的不断增加，5种不同的学习率表现出不一样的预测效果，RMSE值越来越小，呈现较为明显的下降趋势，并慢慢地达到收敛状态。此外，也可看出，虽然随着迭代次数的不断增加，其RMSE值越来越小，即模型预测精度有所提升，但模型训练时间也会不断增加。

4.2 不同学习率对rsrp误差率均值的影响

如图4所示，横坐标表示学习率，纵坐标表示所有

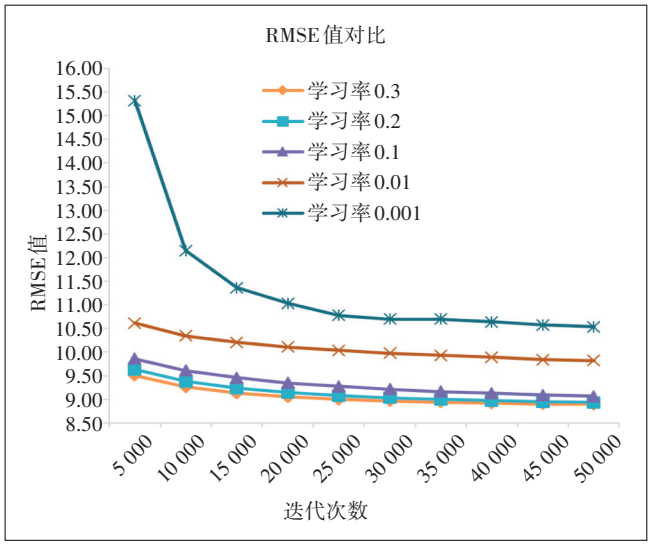

图3　不同学习率的RMSE结果对比

rsrp误差率的平均值。模型迭代次数为50 000次，随之学习率不断减小，其rsrp误差率均值越来越大，最大差值达到5.1%，即反应了模型效果越来越差。从图4中可以看出，学习率为0.3时，rsrp误差率均值达到最小，模型的泛化性更好。

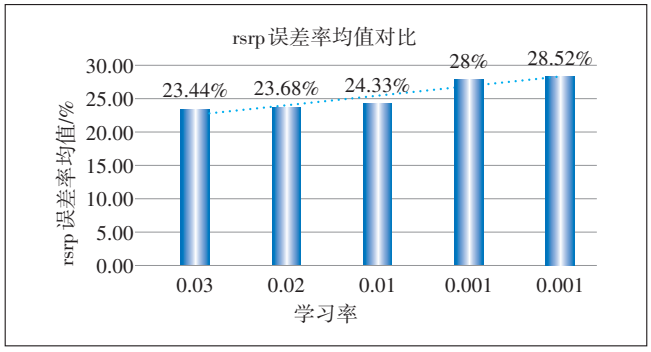

图4　不同学习率的rsrp误差率均值对比

综合图3和图4可知，RMSE值越小，rsrp误差率均值越小，二者相辅相成，并且都能够体现模型对信号预测的效果，RMSE值越小，模型预测精度越高。

4.3 同学习率、不同迭代次数时不同模型RMSE值对比

如图5所示，横坐标表示迭代次数，纵坐标表示RMSE值。从图5可以看出，当学习率为0.3时，随着迭代次数的不断增加，RMSE值越来越小，3个模型均慢慢地达到收敛状态。另外在同样的迭代次数下，LightGBM模型训练结果RMSE值始终优于XGBOOST和GBDT，其平均预测精度要比XGBOOST高出8.4%，比GBDT高出15.36%。

4.4 同迭代次数、不同学习率时不同模型RMSE值对比

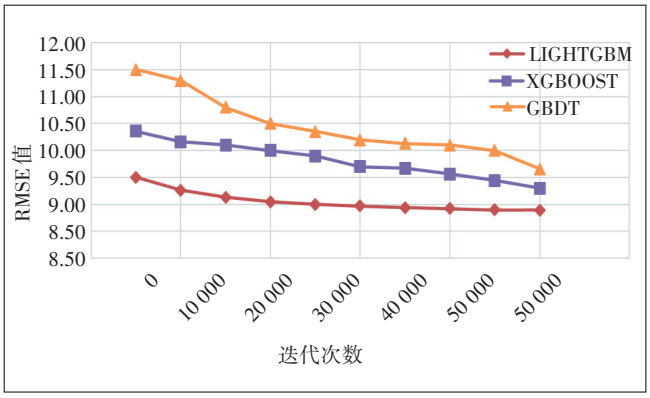

图5　同学习率、不同迭代次数时不同模型RMSE值对比

如图6所示，横坐标表示学习率，纵坐标表示RMSE值。图6反映了当迭代次数为30 000时，随着学习率的不断减小，RMSE值越来越大，说明了3个模型在学习率为0.3时，取得较好效果，在学习率为0.001时，模型效果最差。另外在同样的学习率下，LightGBM模型下训练结果RMSE值始终优于XGBOOST和GBDT，其平均预测精度比XGBOOST高出13.84%，比GBDT高出27.85%。

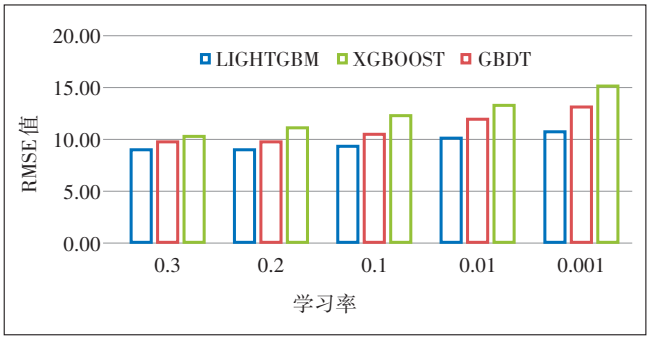

图6　同迭代次数、不同学习率时不同模型RMSE值对比

5 结束语

本文提取了某市各区域基站的MR样本数据，首先对数据进行栅格化并清洗，再对模型进行训练，从网络信号强度预测结果得出：使用LightGBM算法进行预测，修改训练迭代次数和学习率参数，模型训练取得了较好的效果，令人较满意。从rsrp误差率均值可以得出：不同学习率下，随着迭代次数的增加，模型能够快速收敛，且模型训练效果也越来越好。

参考文献：

[1] 罗凡云，郭俊峰. TD-LTE网络覆盖性能分析[J]. 移动通信，2010，34(5)：41-44.
[2] 林世明，高志斌，高凤连，等. 基于路测的TD-LTE网络优化分析[J]. 现代电子技术，2015(9)：20-23.
[3] 何少尉. 基于CQT与DT的通信网优化研究[J]. 科技信息，2012(24)：295-297.
[4] 宫元峰，黄轶. 基于大数据分析的室内深度覆盖优化方法研究[J]. 电信科学，2019，35(05)：149-154.
[5] 匡红，叶猛. LTE中基于S1接口的数据采集系统研究[J]. 电视技术，2013，37(3)：106-108.
[6] 黄剑锋. LTE网络中基于干扰概率进行干扰分析的方法及系统的制作方法：CN105208581A[P]. 2015-12-30.
[7] WITTEN I H，FRANK E . Data Mining：Practical Machine Learning Tools and Techniques (Third Edition)[M]. 机械工业出版社，2005.
[8] 高洁，张涛，程新洲，等. 一种基于LightGBM机器学习算法的用户年龄及性别预测方法[J]. 邮电设计技术，2019(9)：36-39.
[9] FRIEDMAN J H . Greedy Function Approximation：A Gradient Boosting Machine[J]. Annals of Statistics，2001，29(5)：1189-1232.
[10] SON J，JUNG I，PARK K，et al. Tracking-by-Segmentation with Online Gradient Boosting Decision Tree[C]// IEEE International Conference on Computer Vision. IEEE，2016.
[11] SHERIDAN R P，WANG W M，LIAW A，et al. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships[J]. Journal of Chemical Information & Modeling，2016：2353.
[12] SONG R，CHEN S，DENG B，et al. eXtreme Gradient Boosting for Identifying Individual Users Across Different Digital Devices[C]// International Conference on Web-Age Information Management. Springer International Publishing，2016.
[13] WANG D，ZHANG Y，ZHAO Y. LightGBM：An Effective miRNA Classification Method in Breast Cancer Patients[C]// the 2017 International Conference. 2017.
[14] MACARINI L，MERRY W J，PATERNAIN G P . On the growth rate of leaf-wise intersections[J]. Journal of Symplectic Geometry，2011，10(4)：601-653.
[15] 姜备. 基于LTE MR的移动网络优化研究[D]. 上海：上海师范大学，2016.
[16] CHAI T，DRAXLER R R . Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature[J]. Geoscientific Model Development，7，3(2014-06-30)，2014，7(3)：1247-1250.
[17] WILLMOTT C，MATSUURA K . Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance[J]. Climate Research，2005，30(1)：79.

作者简介：
张硕伟，毕业于上海师范大学，工程师，硕士，主要从事机器学习、大数据移动网络分析工作；裴明丽，毕业于安徽大学，工程师，硕士，主要从事算法相关工作；高有利，毕业于东北农业大学，高级工程师，学士，主要从事网络平台和应用的架构设计及项目研发工作；黄铭，毕业于上海海事大学，工程师，硕士，主要从事机器学习、网络服务智能化研发工作；刘贤松，毕业于武汉水利电力大学(武汉大学)，中国联通网络AI中心副经理，硕士，主要分管网络服务智能化产品研发和网络服务智能化产品推广管理工作。