A XGBoost Based Wireless Interference Relation Mining and Performance Prediction Method

Han Liu, Tao Peng, Yichen Guo, Yachen Wang, Gonglong Chen, Feng Yang, Wei Chen
Wireless Signal Processing and Networks Laboratory (WSPN)
Key Laboratory of Universal Wireless Communication, Ministry of Education
Beijing University of Posts and Telecommunications (BUPT), Beijing, 100876, P.R. China
Email: {liuhanhan, pengtao, guoyichen}@bupt.edu.cn, 1475324023@qq.com,
secpros@163.com, yangfengee04@gmail.com, chenweicmri@139.com

2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall) | 978-1-6654-1368-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/VTC2021-FALL52928.2021.9625052

***Abstract*—Ultra-dense network (UDN) is considered to be the key technology for the fifth generation (5G) networks to provide high capacity. However, intensive deployment of femtocells bring severe inter-cells interference (ICI), which greatly limits the performance of the network and the capacity gain the system can obtain. Therefore, the key to solve this problem is to obtain accurate interference information through accurate interference modeling. In fact, the wireless big data generated during the operation of the wireless network contains rich wireless interference information. Based on this, this paper proposes an uplink interference identification and signal-to-interference-plus-noise ratio (SINR) prediction algorithm based on XGBoost and interference model. The proposed algorithm uses the wireless big data generated during network operation to train the XGBoost algorithm, mining the signal-to-interference ratio (SIR) and signal-to-noise ratio (SNR) information between links in the wireless network without increasing the overhead of wireless resources, and then combining with the proposed interference model to achieve accurate prediction of the SINR. The simulation results show that when the training data of the target user reaches 5000 pieces, the prediction error of its SINR will be reduced to less than 0.5dB, which effectively reduces the requirement of data quantity and computing power, and can meet the practical application requirements.**

***Index Terms*—Ultra-dense network, machine learning, XGBoost, big data**

I. Introduction

To cope with the rapid increasing demand for wireless network, ultra-dense network (UDN) is a promising solution [1] and it had already become one of the key networking mode in the fifth generation (5G) network. Network densification increases the density of access points, tremendously improve the capacity of the network. However, the densification will inevitably induce severe inter-cell interference (ICI). Although this further improves the performance of wireless access network, it also makes the inter-cell interference in 5G network show a high degree of uncertainty, dynamics, severity and other characteristics that have never been shown in previous networks, which brings severe challenges to network interference control. There are mainly three interference coordination methods in traditional wireless networks: inter cell interference coordination (ICIC) and enhanced ICIC (eICIC), coordinated multiple points (CoMP), self-organizing network(SON). ICIC and eICIC coordinate inter-cell interference through signal interaction of interference indication information between base stations (BSs). This can cause very high signaling exchange overhead, and interference information are relatively simple and outdated. CoMP channel estimation needs to use the reference signal for channel estimation to obtain accurate channel state information from each user terminal to all cooperative BSs. The resulting channel estimation error will seriously restrict the improvement of system performance. Due to its own positioning and capabilities, SON technology mainly deals with network changes with large time granularity (usually in hours), and is unable to deal with highly dynamic and complex interference of 5G network in small time scales.

To mitigate these problems, some works have been done. The distribution density of the BSs is considered in reference [2]. However, the density of BSs doesn't directly relate to the interference relationship. The same problem exists in [3], [4] and [5], which use the distance between BSs to estimate interference. Reference [6] requires user equipments (UEs) to identify the interference BSs, which may be burdensome to some UEs. The aforementioned works [2]- [6] are focused on downlink interference. In [7], uplink communication is taken into account, while the interference calculation requires channel gain from the node and the packet head, which is difficult to obtain. All of the above work has some unrealistic assumptions due to inadequate use of information or incomplete access to data. To obtain global information, a centralized wireless access network (C-RAN) architecture is introduced. Reference [8] directly use the measured signal-to-interference-plus-noise ratio (SINR) to define the interference relationship between two UEs connected to different BSs. However, it is likely that multiple UEs share the same resource block (RB), so this approach may not be accurate enough to eliminate the impact of additional sources of interference.

To dig further into huge amount of data collected in C-RAN based network, machine learning (ML) algorithms are ideal choices. Our previous work [9] and [10] proposed multilayer perceptron algorithm. The algorithm could identify different uplink interference sources and their strengths for each user. And [11] proposed an algorithm based on association rules, which could sort downlink interference sources accurately. In

This work was supported in part by the Beijing Natural Science Foundation (No. L192002) and the National Natural Science Foundation of China (No. 61631004).

addition, on the basis of in-depth analysis of the mechanism of ICI in wireless networks, a SINR prediction algorithm based on regression is proposed in [12]. These algorithms achieve great precision with the need of only data collected during the normal operation of network, which shows great potential of ML algorithms in this field.

Inspired by [9] and [12], in this paper, an explainable, data-driven and XGBoost based uplink interference identification and SINR prediction algorithm is proposed. Other than conventional methods introduced above, the proposed approach provides a complete and accurate uplink user interference modeling scheme by collecting large-scale RB allocation data and corresponding up-link SINR data, and using big data analysis and ML algorithm to mine the wireless network interference relationship between users. Through the analysis of the interference mechanism, the algorithm is proved to be highly explanatory and effective, and the accuracy of interference recognition and prediction is extremely high.

The rest of this article is organized as follows. Section II introduces the system model of this paper. Section III elaborates the principle and derivation of the proposed algorithm in detail. In Section IV, we design a simulation to evaluate the proposed algorithm and present the results and analysis. Finally, the conclusion is drawn in Section V.

II. System Model

In this paper, up-link communications in UDNs especially the condition of stadiums, concerts and shopping malls outlined as "Great service in a crowd" [13], are considered. In these scenarios, the ICI is particularly severe and the interference management is extremely imperative.

In Fig. 1, a dual-strip UDN model is illustrated, with large amounts of femtocell access points (FAPs) deployed using C-RAN architecture. The centralized feature of the C-RAN makes it appropriate to support cooperative techniques such as interference management and facilitate 5G technologies such as UDNs. Further, there are J subareas in this area and each subarea contains only one FAP, so all FAPs could be denoted as set $\mathbf{F} = \{FAP_1, FAP_2, \ldots, FAP_j, \ldots, FAP_J\}$. And set $\mathbf{U} = \{UE_1, UE_2, \ldots, UE_n, \ldots, UE_N\}$ consists of all active N user equipments (UEs).

The algorithm proposed in this paper is actually applicable to any wireless communication networking model, and its goals may vary according to the different key technologies used in the wireless system. In the orthogonal resource sharing system among different users in the cell, such as orthogonal frequency-division multiple access (OFDMA) and time division multiple address (TDMA) system, there is no interference between users in the cell, only mining the user interference between different cells. However, for the system in which users still have interference in the cell, such as code division multiple access (CDMA) system, the interference between different users in the same cell and between cells can be mined and identified accurately. Without loss of generality, this paper takes the OFDMA wireless system with UDN as a typical system scenario to carry out the subsequent technical

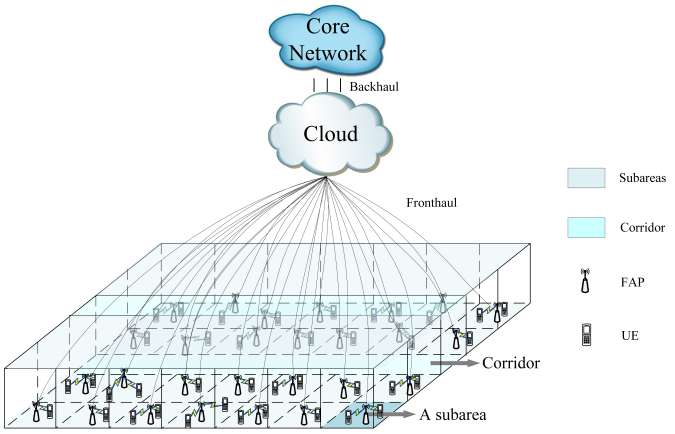

Fig. 1. System model of a C-RAN based UDN [9].

details, which means each RB could only be allocated to at most one user in each cell concurrently. Therefore, users who access to the same FAP do not generate interference to each other. Thus, intra-cell co-channel interference is eliminated. Each FAP could utilize all I RBs in wireless resource set $\mathbf{R} = \{RB_1, RB_2, \ldots, RB_i, \ldots, RB_I\}$ to serve associated users.

Power control and multi input multi output (MIMO) are also considered. The partial path loss compensation is adopted, therefore, the actual transmission power of arbitrary UE_n on each RB in the unit of dB is

$$P_n(dB) = P_0 + \alpha PL_{n,j_n} + 10\lg(n_O) \tag{1}$$

where P_0 is the semi-persistent criterion in the unit of dB, α is the path loss compensation factor, PL_{n,j_n} is the path loss between UE_n and its associated FAP_{j_n}, j_n is the index of UE_n's associated FAP, and n_O is the number of output antennas of any UE. Besides, the total transmission power of a UE on all allocated RBs cannot exceed the maximum UE transmission power P_{max} which is fixed and works for all UEs.

III. Uplink Interference Identification and SINR Prediction Based on XGBoost

A. Data Collection and Collation

In this paper, resource allocation configurations for a large number of RBs and the corresponding measured SINRs are collected from the network as training data. These data are continuously collected and stored in the database, and converted into the format showed in Fig. 2 before training by XGBoost. Each user has its own dataset and the inter-user interference relation of each user is learned based on its dataset respectively because the interfering user set of each UE is different. Each entry in the data set shows one RB resource allocating configuration and the corresponding measured uplink SINR.

Dataset of user UE_1 showed in Fig. 2 contains a large number of samples which consist of attributes (i.e., interfering users) and a label (i.e., the up-link SINR on the corresponding

Data set of UE 1		Attribute 1	Attribute 2	---	Attribute |U|-2	Label
		UE 3	UE 4	---	UE |U|	Up-Link SINR(dB)
TTI 1	RB 1	1	1	---	0	3.25
	RB 2	1	0	---	0	4.07
	RB 3	0	1	---	1	13.21
	RB 4	1	0	---	1	3.84
TTI 2	RB 3	1	0	---	0	5.24
	RB 4	1	0	---	1	4.6
	RB 5	1	1	---	1	3.19
	RB 6	0	1	---	0	6.36
	RB 7	1	0	---	0	5.75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
TTI t	RB 9	0	0	---	1	17.34
	RB 10	0	1	---	0	14.85
	RB 11	0	0	---	1	16.47

Fig. 2. Training data set for UE_1.

RB allocated to UE_1). Each line in the yellow area of dataset showed in Fig. 2 is a sample. For example, $\{1, 1, \ldots, 0, 3.25\}$ is the first sample in the dataset of UE_1 corresponding to TTI 1 and RB 1. The attribute value "1" means the specific user is reusing RB 1 with UE_1 in TTI 1, and vise versa for zeros in the entry. And the SINR received by UE_1's associated FAP in TTI 1 on RB 1 is 3.25 dB. The number of samples corresponding to a certain TTI equals to the number of RBs allocated to UE_1 in that TTI. The total number of samples in dataset of UE_1 is the accumulative number of RBs allocated to UE_1 in each up-link TTI.

B. XGBoost Algorithm Introduction

In this paper, the machine learning algorithm chooses to use XGBoost algorithm. XGBoost is short for "Extreme Gradient Boosting", which is derived from the Gradient Boosting framework but is much more efficient. The algorithm is capable of parallel computation, approximate tree construction, efficient processing of sparse data and optimization of memory usage, as well as multiple tasks such as regression, classification and sorting. Due to its strong prediction performance and fast training speed, XGBoost algorithm is selected in this paper to mine and predict interference relations.

The XGBoost algorithm is described in detail below:

1) Prediction model:

The XGBoost tree integration model uses the summation of the output values of K trees to predict the output, and each tree fits the residual of the previous model,

$$\hat{\gamma}=\sum_{k=1}^{K} f_k\left(\mathbf{w}\right) \tag{2}$$

Where $\mathbf{w} = \{w_n\}$ is the resource allocation vector; $\hat{\gamma}$ is the predicted value of SINR; $f_k(\mathbf{w})$ is the weight of the leaf node corresponding to the vector $\mathbf{w}$ in the k tree. Fig. 3 is a schematic of a simple XGBoost prediction model consisting of two trees.

2) The objective function of XGBoost is:

$$L=\sum_{i=1}^{I} l\left(\hat{\gamma}_{(i)}, \gamma_{(i)}\right)+\sum_{k=1}^{n_{\max}} \Omega\left(f_k\right) \tag{3}$$

The first part is the loss function and the second part is the regularization term.

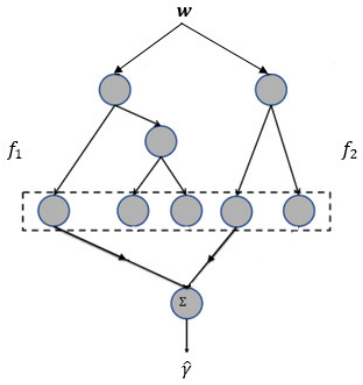

Fig. 3. Schematic of XGBoost prediction model.

The loss function measures the difference between the predicted value and the target value of all I samples. $\hat{\gamma}_{(i)}$ and γ_i represent the predicted value and the target value of the ith sample respectively. Here, we use the root mean square error (RMSE):

$$\sum_{i=1}^{I} l\left(\hat{\gamma}_{(i)}, \gamma_{(i)}\right)=\sqrt{\frac{1}{I} \sum_{i=1}^{I}\left(\hat{\gamma}_i-\gamma_i\right)^2} \tag{4}$$

The regularization term is used to control the complexity of the tree to prevent overfitting:

$$\Omega\left(f\right)=\alpha T+\frac{1}{2} \beta \sum_{t=1}^{T} s_t^2 \tag{5}$$

Where T is the total number of leaf nodes of tree f; s_t is the weight of the t leaf node of tree f; α and β are hyperparameters to determine the regularization intensity of L0 and L2 of tree f, respectively. The larger these hyperparameters are, the simpler the tree structure is expected to be.

3) Determine the weight of leaf node of the subtree k:

The characteristic of XGBoost algorithm is the second-order Taylor approximation expansion of the loss function, so the loss function can be customized as long as the second-order differentiability is satisfied:

$$l\left(\hat{\gamma}_{(i)}, \gamma_{(i)}\right) \approx g_i f_k\left(\mathbf{w}_{(i)}\right)+\frac{1}{2} h_i f_k^2\left(\mathbf{w}_{(i)}\right)+\text{constant} \tag{6}$$

The first order gradient g_i and second order gradient h_i of each sample i are calculated. Then substitute the objective function L, and the sample set in leaf node t is D_t, then:

$$L \approx \sum_{t=1}^{T}\left[\left(\sum_{i \in D_t} g_i\right) s_t+\frac{1}{2}\left(\sum_{i \in D_t} h_i+\beta\right) s_t^2\right]+\alpha T \tag{7}$$

The above equation is a quadratic function of the weight s_t of leaf node t; let $G_t = \sum_{i \in D_t} g_i$, $H_t = \sum_{i \in D_t} h_i + \beta$, easily obtain the optimal weight:

$$s_t^*=-\frac{G_t}{H_t+\beta} \tag{8}$$

4) The goal of XGBoost algorithm training is to minimize the objective function L:

First, for each subtree, the leaf nodes are constantly split and the node weights are updated. The original leaf node t_S is split into two new leaf nodes, t_L and t_R, and the change of the objective function is as follows:

$$\Delta L = \frac{1}{2}\left[\frac{G_{t_S}^2}{H_{t_S}+\beta} - (\frac{G_{t_L}^2}{H_{t_L}+\beta} + \frac{G_{t_R}^2}{H_{t_R}+\beta})\right] + \gamma \quad (9)$$

Where, the first term is the fraction value of the parent node before segmentation, the second and third terms are the sum of the fractions of the left and right subtrees after segmentation, and the last term is the complexity caused by the introduction of additional leaf nodes.

There may be multiple samples in the original leaf node t_S, so there are multiple splitting schemes, corresponding to different ΔL. Therefore, the scheme with ΔL minimum and less than 0 was selected for node splitting and weight updating. Until at least one of the following conditions is satisfied for each leaf node: the sample number contained in the leaf node is not greater than the minimum sample number ε; the depth of leaf nodes reaches the maximum tree depth $d_{\max}$; all the splitting scheme for the leaf node, $\Delta L \geq 0$.

When the leaf nodes of the subtree stop splitting, the subtree is created. Then, create a new subtree and split it. The previously created subtree will not change. Until the number of subtrees reaches the maximum number of subtrees K, the creation of subtrees stops and the training process ends.

C. Interference Model

Based on the typical system model and scene in Section II above, the new technical scheme proposed in this paper will deeply analyze the generation mechanism of inter-cell interference in wireless network. And on this basis, the SINR prediction technology based on wireless interference model is proposed.

In the uplink direction of wireless network, the signal power of UE_m received by FAP_{j_m} is

$$S_m^{j_m} = P_m G_m^{j_m} \quad (10)$$

where $G_m^{j_m}$ is the channel gain from UE_m to FAP_{j_m}.

The interference signal power from UE_n, which occupies the same wireless resources as UE_m, to FAP_{j_m} is:

$$I_n^{j_m} = P_n G_n^{j_m} \quad (11)$$

Where $G_n^{j_m}$ is the channel gain from the interference user UE_n to FAP_{j_m}.

SINR of UE_m measured in FAP_{j_m} could be expressed as

$$\gamma_m = \frac{S_m^{j_m}}{\sum_{n\in\mathbf{N}\setminus\mathbf{N_m}} w_n I_n^{j_m} + \sigma^2} \quad (12)$$

where $\mathbf{N}$ and $\mathbf{N}_{j_m}$ are the user set in the system, and the user set without interference with user m respectively. And $w_n \in \{0,1\}$ is the resource allocation indicator. $w_n = 1$ means U_n is sharing the same RB with U_m, while $w_n = 0$ indicates otherwise. σ^2 is power of additive noise.

Further transformation of the above equation can be obtained:

$$\begin{aligned}\frac{1}{\gamma_m} &= \frac{\sum_{n\in\mathbf{N}\setminus\mathbf{N_m}} w_n I_n^{j_m} + \sigma^2}{S_m^{j_m}} \\ &= \sum_{n\in\mathbf{N}\setminus\mathbf{N_m}} \frac{1}{w_n \frac{S_m^{j_m}}{I_n^{j_m}}} + \frac{1}{\frac{S_m^{j_m}}{\sigma^2}} \\ &= \sum_{n\in\mathbf{N}\setminus\mathbf{N_m}} w_n \frac{1}{\gamma_{m,n}} + \frac{1}{\gamma_{m,\sigma^2}}\end{aligned} \quad (13)$$

Where, $\gamma_{m,n} = \frac{S_m^{j_m}}{I_n^{j_m}}$is the signal-to-interference ratio (SIR) of UE_m to UE_n; $\gamma_{m,\sigma^2} = \frac{S_m^{j_m}}{\sigma^2}$ is the signal-to-noise ratio (SNR) of the received signal of UE_m (without considering interference).

Let $\widetilde{\gamma}_m = \frac{1}{\gamma_m}$, $\widetilde{\gamma}_{m,n} = \frac{1}{\gamma_{m,n}}$, $\widetilde{\gamma}_{m,\sigma^2} = \frac{1}{\gamma_{m,\sigma^2}}$, then the above equation can be simplified as:

$$\widetilde{\gamma}_m = \sum_{n\in\mathbf{N}\setminus\mathbf{N_m}} w_n \widetilde{\gamma}_{m,n} + \widetilde{\gamma}_{m,\sigma^2} \quad (14)$$

Then, considering that UL-SINR is usually used in dB in actual usage scenarios, γ_m value is converted to dB. That is, the logarithm of both sides of the equation is taken to obtain the following formula:

$$\gamma_m = -10\lg\widetilde{\gamma}_m = -10\lg\left(\sum_{n\in\mathbf{N}\setminus\mathbf{N_m}} w_n \widetilde{\gamma}_{m,n} + \widetilde{\gamma}_{m,\sigma^2}\right) \quad (15)$$

D. Online Prediction

The interference prediction framework based on XGBoost (XGB-PF) is shown in Fig. 4, which mainly includes two parts: offline prediction model training and online prediction of interference intensity. In this paper, XGBoost algorithm is used for off-line training to mine the interference relationship, and network interference relationship modeling function is used for online interference prediction.

1) Interference relation ($\widetilde{\gamma}_{m,n}$, $\widetilde{\gamma}_{m,\sigma^2}$) mining:

Firstly, a large amount of data as shown in Fig. 2 is used to conduct off-line training on XGBoost to establish the interference model. Then, XGBoost uses the trained interference model to mine the interference relationship of UE_m $\widetilde{\gamma}_{m,n}$, $\widetilde{\gamma}_{m,\sigma^2}$. At this time, the input data format is shown in Fig. 5. The corresponding input data of no interference user and all different single interference user UE_n are input into the prediction model respectively, and the corresponding SNR_{m,σ^2} and $SINR_{m,n}$ are obtained from the output, the unit is dB. Finally, convert the dB value to the true value, through $\widetilde{\gamma}_{m,\sigma^2} = -10\lg\left(SNR_{m,\sigma^2}\right)$ and $\widetilde{\gamma}_{m,n} = -10\lg\left(SINR_{m,n}\right) - \widetilde{\gamma}_{m,\sigma^2}$ to calculate $\widetilde{\gamma}_{m,n}$, $\widetilde{\gamma}_{m,\sigma^2}$, then the corresponding interference relation is extracted.

In Fig. 5, each feature attribute represents other UE in the system who might interfere with the UE_m. Each row of data represents the interfering user of a certain RB in a certain TTI for UE_m. "1" indicates that the interfering user multiplexed the RB with UE_m, and "0" indicates that the interfering user did not multiplexed the RB. Therefore, each data sample represents that a TTI UE_m shares a RB with

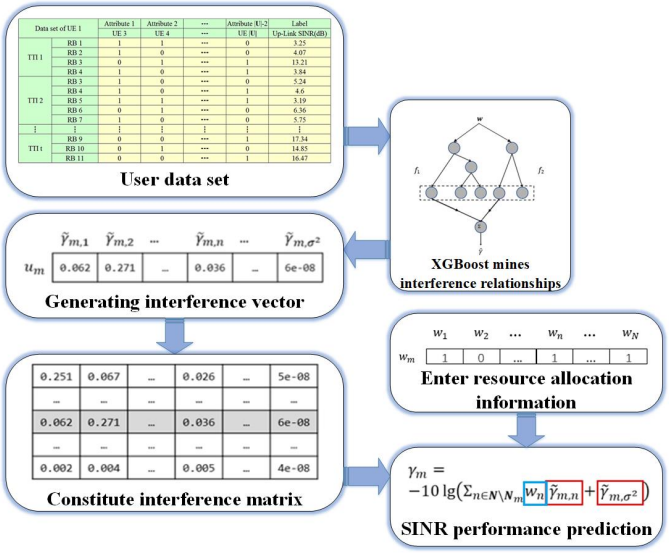

Fig. 4. XGB-PF interference prediction framework.

UE_m	UE_1	UE_2	⋯	UE_{m-1}	UE_{m+1}	⋯	UE_N	UL-SINR（dB）
RB	0	0	⋯	0	0	⋯	0	Predicted value of UE_m's own SNR_{m,σ^2}
RB	1	0	⋯	0	0	⋯	0	Predicted value of $SINR_{m,1}$ which UE_m and UE_1 shared resources
RB	0	1	⋯	0	0	⋯	0	Predicted value of $SINR_{m,2}$ which UE_m and UE_2 shared resources
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
RB	0	0	⋯	0	0	⋯	1	Predicted value of $SINR_{m,N}$ which UE_m and UE_N shared resources

Fig. 5. Data sets used to extract $\widetilde{\gamma}_{m,n}$ and $\widetilde{\gamma}_{m,\sigma^2}$.

a user, and the SINR when UE_m interferes with any other single user can be obtained by input this data sample into the prediction model.

2) SINR (γ_m) prediction of interference performance:

After the corresponding interference relation $\widetilde{\gamma}_{m,n}$ and $\widetilde{\gamma}_{m,\sigma^2}$ are mined through XGBoost, the network interference relation modeling function (15) deduced in Section III-C is used for interference prediction: $\gamma_m = -10\lg\widetilde{\gamma}_m = -10\lg\left(\sum_{n\in\mathbf{N}\setminus\mathbf{N_m}} w_n\widetilde{\gamma}_{m,n} + \widetilde{\gamma}_{m,\sigma^2}\right)$.

Through this function, the UL-SINR in the case of arbitrary multiple interference users can be predicted by inputting the corresponding interference vector $\mathbf{w}$ with w_n equal to 0 or 1. In the subsequent test case set, the interference vector data corresponding to different multi-interference users are randomly input into the prediction model, and the corresponding interference intensity is predicted, which is compared with the theoretical value of SINR, and RMSE is used to represent the prediction performance.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Side length of the square femtocells	10 m
Width of the corridor	20 m
femtocell number per row	4
FAP number per femtocell	1
Minimum distance between FAPs	8 m
UE number per femtocell q	2
System bandwidth W	5 MHz
Semi-persistent criterion P_0	-67 dB
Max transimssion power of UE P_{max}	23 dBm
Path loss compensation factor α	0.7
White noise power density	-174 dBm/Hz
Path loss model	$(38.46 + 20 \times \lg d)$ dB
Antenna configuation$n_I \times n_O$	2×2
The test set size M_t	20000
The maximum depth of the tree $d_{\max}$	30
Minimum sample size ε	1
L0 regularization parameter α	5
L2 regularization parameter β	30
Learning rate η	0.1
The number of trees K	100

IV. SIMULATION RESULTS AND ANALYSIS

A. Simulation Platform

All the data used for training and verifying algorithms in this paper were collected from a system-level open source long term evolution (LTE) network dynamic simulation platform, i.e., LTE-Sim [14]. LTE-Sim is a system-level open source framework to simulate LTE networks developed by Polytechnic University of Bari.

B. Simulation Parameters

As illustrated in Fig. 1, a dual-strip UDN model is created. With 4 femtocells per row and 4 rows in total, there're 16 femtocells in the model ($J = 16$). System bandwidth is configured as 5 MHz, consisting with 25 RBs. Path loss is the same as that defined in [15]:

$$PL_{n,j}(dB) = 38.46 + 20\lg d_{n,j} \tag{16}$$

where $d_{n,j}$ is the distance between U_n and SBS_j, and $PL_{n,j}$ is expressed in decibel. The number of input antennas for every FAP is $n_I = 2$, and $n_O = 2$ as well. In addition, $P_{max} =$ 23dBm (200mW), and the power density of additive Gaussian white noise (AWGN) is -174dBm/Hz.

The comprehensive list of simulation parameters are shown in Table I.

XGBoost parameter settings are optimized as follows: the maximum depth d_max of the algorithm input tree can be set to the number of UE that may interfere with UE_m. It is verified that when the value is less than this value, the performance is not optimal; when the value is greater than this value, the performance does not improve, and the training cost becomes large. In addition, each leaf node needs at least 1 sample to estimate, so the minimum sample number in leaf nodes can be set to 1; The rest of the parameters can be tried and adjusted according to the advantages and disadvantages of RMSE.

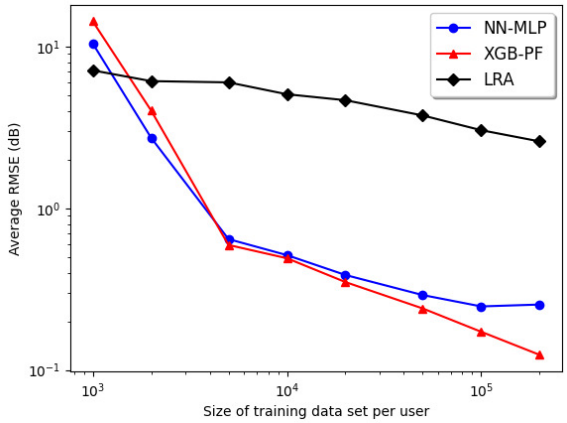

Fig. 6. Interference source identification performance.

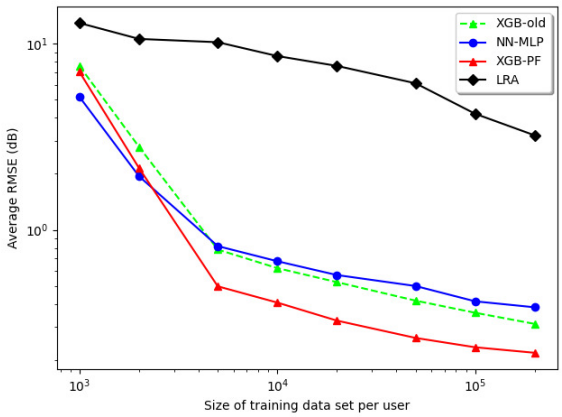

Fig. 7. SINR predicts performance.

C. Simulation Results

According to the XGBoost algorithm proposed in this paper, the performance of interference source identification performance, SINR prediction performance, algorithm training time and the time to achieve the same performance are evaluated and shown, and the performance of the neural network – multilayer perceptron (NN-MLP) algorithm and linear regression algorithm (LRA) are compared.

1) Interference source identification performance (SIR):

Interference source identification performance as shown in Fig. 6, as the amount of training data per user increases, the performance of XGB-PF algorithm quickly outperforms LRA and NN-MLP algorithm. When the number of samples reaches 2000, the performance of XGB-PF algorithm is better than that of LRA. When the number of samples reaches 5000 samples, the performance of XGB-PF algorithm is better than that of NN-MLP algorithm. When the data volume of each user reaches 200,000 samples, the RMSE prediction error of XGB-PF algorithm has reached 0.12dB, which is far better than the NN-MLP algorithm's 0.25dB and the LRA's 2.59dB.

Because accurate identification of interference source can improve the performance of traditional technologies such as ICIC, eICIC and CoMP, the performance of XGB-PF algorithm is far better than that of NN-MLP algorithm and LRA. Therefore, the XGB-PF algorithm can be used to assist ICIC, eICIC and CoMP to improve their performance.

2) SINR predicts performance:

Since the direct use of XGBoost algorithm can also make SINR prediction and obtain results, XGB-old method is added in Fig. 7: mining interference relationship and predicting SINR performance directly through XGBoost algorithm, that is, XGBoost algorithm directly identifies interference source and predicting SINR performance after training. It can be seen that although the performance of this method is slightly better than that of the NN-MLP algorithm, it is worse than that of the XGB-PF algorithm proposed in this paper, so it is not adopted. Meanwhile, since both XGB-PF and XGB-old algorithms directly use XGBoost algorithm for interference source identification and have the same recognition performance, this method is not shown in Section IV-C1.

For each user, when 5000 samples (second-level data) are used for training, XGB-PF algorithm can meet the prediction performance requirement of SINR prediction performance error less than 0.5dB, reaching 0.49dB. It is far better than the LRA (2.78dB) and NN-MLP algorithm (0.81dB). This proves that the XGB-PF algorithm can effectively reduce the requirement of data quantity.

We also found that when the prediction performance of SINR of NN-MLP algorithm reaches 0.5 dB, 50000 samples are needed, which is 10 times that of XGB-PF algorithm. LRA convergence is too slow, even when 200,000 samples, the prediction performance is still greater than 1 dB, the performance is too poor.

In fact, we also did some research on scalability, such as three UEs per cell and ten UEs per cell, for a total of 48 and 160 users. After simulation verification, the convergence of SINR prediction by XGBoost is still good when the base station users increase, and the convergence trend is similar to that of 2 users. With the increase of users, the prediction accuracy decreases slightly, but the accuracy is always better than that of the neural network scheme.

3) Algorithm training time:

As shown in Fig. 8, the time consuming of XGB-PF algorithm is one order of magnitude less than that of NN-MLP algorithm algorithm under the condition of the same size of training data set, which is basically the same as that of XGB-old method. When the prediction error of SINR is less than 0.5dB, that is, the training sample is 5000, the training time required by each user of XGB-PF algorithm is less than 1s and reaches 0.34s, and each user can achieve sub-second training. However, when the NN-MLP algorithm has the same training data set of 5000, each user needs about 4.26s training time, which cannot well meet the requirements. LRA has the shortest training time, but the prediction performance of SINR is too poor. This also proves that the XGB-PF algorithm can effectively reduce the computing power requirements.

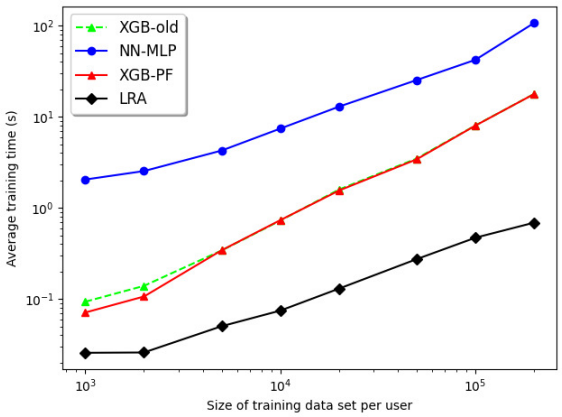

Fig. 8. Algorithm training time.

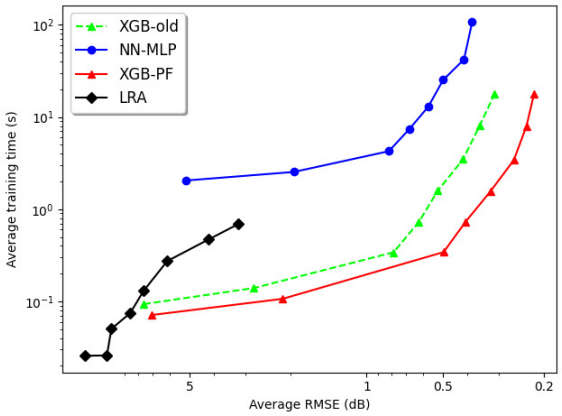

Fig. 9. Training time to achieve the same performance.

4) Training time to achieve the same performance:

As shown in Fig. 9, the training time of the NN-MLP algorithm is two orders of magnitude longer than that of the XGB-PF algorithm under the condition that the same prediction accuracy is 0.5dB. Compared with the LRA, the XGB-PF algorithm can reduce the average prediction error by an order of magnitude under the same time consumption, and the improvement is also very significant.For the performance shown inFig. 9, the ordinates of Fig. 7 and Fig. 8 are taken as the abscissa and ordinate of Fig. 9 respectively to make a more intuitive display. Therefore, the size of the training set is the independent variable.

V. Conclusion

In this paper, an interference prediction method based on XGBoost machine learning algorithm is proposed. This algorithm utilizes a large amount of wireless resource allocation data and wireless measurement data generated in the scheduling process, and mining the wireless network interference relationship between users through big data analysis and machine learning algorithm. At the same time, through in-depth analysis of the interference mechanism and relevant information in the wireless network, the wireless network interference model formula is deduced, and combined with XGBoost algorithm to predict the interference performance, so as to provide a complete and accurate uplink interference modeling scheme. The simulation results show that, compared with the disadvantages of the long learning time of NN-MLP algorithm and the poor performance of LRA, XGBoost algorithm can achieve the amount of second-level training data and subsecond-level training time when the predicted performance meets the requirements. Therefore, the scheme in this paper is simple to implement, closer to the actual network scene, and realizes real-time, efficient, high-precision and complete interference prediction. It will also facilitate the subsequent resource allocation process.

References

[1] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2134–2168, 2019.

[2] W. Li, J. Zhang, "Cluster-based resource allocation scheme with QoS guarantee in ultra-dense networks," *IET Commun.*, vol. 12, issue 7, pp. 861–867, Apr 2018.

[3] C. Niu, Y. Li, R. Q. Hu, F. Ye, "Fast and efficient radio resource allocation in dynamic ultra-dense heterogeneous networks," *IEEE Access*, vol. 5, pp. 1911–1924, Feb 2017.

[4] Y. Wang, Z. Tan, "Graph-based and QoS guaranteed spectrum allocation for dense local area femtocell networks," in *Proc. IEEE Military Communications Conference*, Oct 2014, pp. 1556–1561.

[5] Y. Ye, H. Zhang, X. Xiong, Y. Liu, "Dynamic min-cut clustering for energy savings in ultra-dense networks," in *Proc. IEEE VTC*, Sept 2015, pp. 1–5.

[6] W. Jiang, P. Hong, K. Xue, Q. Fan, T. Hu, "QoS-aware dynamic spectrum resource allocation scheme in C-RAN based dense femtocell networks," in *Proc. WCSP*, Oct 2015, pp. 1–6.

[7] S. Basloom, A. Nazar, G. Aldabbagh, M. Abdullah, N. Dimitriou, "Resource allocation using graph coloring for dense cellular networks," in *Proc. ICNC*, Feb 2016, pp. 1–5.

[8] C. Zhao, X. Xu, Z. Gao, L. Huang, "A coloring-based cluster resource allocation for ultra dense network," in *Proc. IEEE ICSPCC*, Aug 2016, pp. 1–5.

[9] J. Cao, X. Liu, W. Dong, T. Peng, R. Duan, Y. Yuan, W. Wang, "A neural network based conflict-graph construction approach for ultra-dense networks," in *Proc. IEEE GLOBECOM Workshops*, Dec 2018, pp. 1–6.

[10] J. Cao, T. Peng, X. Liu, W. Dong, R. Duan, Y. Yuan, W. Wang, S. Cui, "Resource allocation for ultra-dense networks with machine learning based interference graph construction," *IEEE Internet Things J.*, in press.

[11] J. Cao, T. Peng, W. Dong, X. Liu, W. Wang, "An association rules based conflict-graph construction approach for ultra-dense networks," in *Proc. IEEE GLOBECOM Workshops*, Dec 2018, pp. 1–7.

[12] Y. Guo, C. Hu, T. Peng, H. Wang and X. Guo, "Regression-Based Uplink Interference Identification and SINR Prediction for 5G Ultra-Dense Network," in *Proc. IEEE ICC*, 2020, pp. 1-6

[13] METIS, "The 5G future scenarios identified by METIS - The first step toward a 5G mobile and wireless communications system," https://www.metis2020.com/press-events/press/the-5gfuture-scenarios-identified-by-metis/.

[14] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, P. Camarda, "Simulating LTE cellular systems: An open source framework," *IEEE Trans. Veh. Tech.*, vol. 60, no. 2, pp.498–513, Feb 2011.

[15] S. Abeta, "3GPP TR 36.814: Further advancements for E-UTRA physical layer aspects," the 3rd Generation Partnership Project (3GPP), 2017.