Load Balancing and User Association Based on Historical Data

Yuejie Zhang[1], Kai Sun[1], Xueliang Gao[1], Wei Huang[1], Haijun Zhang[2]
[1] *College of Electronic Information Engineering, Inner Mongolia University, Hohhot, China*
[2] *University of Science and Technology Beijing*, Beijing, China
Emails: sunkai@imu.edu.cn, huangwei@imu.edu.cn

GLOBECOM 2021 - 2021 IEEE Global Communications Conference | 978-1-7281-8104-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/GLOBECOM46510.2021.9685782

***Abstract*—With the rapid increase of demand on mobile data traffic of user equipment (UE), network operators have begun to deploy abundant heterogeneous base stations (BSs) to ensure the quality of service (QoS) of UEs, which will cause new problems such as network congestion and load imbalance. If the pattern of user association (UA) can be adjusted in accordance with the results of traffic prediction, the performance of system will be greatly improved. Therefore, a new neural network approach based on spatial and temporal characteristics of traffic data is proposed for traffic prediction. The fluctuations of traffic in the future week are predicted by the proposed method. Then, UA is represented as a problem of maximizing the utility function of load balancing index, and a dynamic user association based on load prediction algorithm (DUALP) which aims to achieve a proactive load balancing is proposed. The QoS of UEs is ensured and the long-term stability of the system is achieved by DUALP. Experimental results show that compared to the classic UA strategies, the most optimal load distribution is realized by DUALP.**

***Index Terms*—user association, ultra-dense network, load balance, traffic prediction**

I. Introduction

With the large-scale popularization of mobile user equipments (UEs), the mobile traffic data generated at UE side are increasing exponentially [1]. The explosive growth of data traffic has prompted network operators to enhance spectrum reuse by a high-density deployment of base stations (BSs) [2]. Although network densification significantly increases system capacity, it brings new problems and challenges to the communication network. It will result in a shrinking of coverage area and a greater temporal and/or spatial changes of traffic, that is it could lead to load imbalance or the degradation of system performance [3]. Moreover, the user association (UA) based on the received power, by which UE will select BS with the strongest signal to provide service for itself, is difficult to apply to ultra-dense networks (UDN) as it will also cause or aggravate the problem of load imbalance. Considering the above elements, it is necessary to design a UA scheme in accordance with the characteristics of UDN [4].

This work was supported in part by the National Natural Science Foundation of China (Nos. 61861034, 62161035), and in part by the Natural Science Foundation of Inner Mongolia Autonomous Region of China (No. 2019MS06009).

A. Related Works

UA is one of the important ways to achieve load balance among different heterogeneous BSs. The biasing technique is widely used to solve the problem of load imbalance and poor performance of UEs [5], [6]. This method that set a bias value for the received power of low-power BSs can offload UEs to lightly loaded BSs [5], but it is a heuristic-based scheme and it is hard to determine the optimal biasing value [6]. UA methods for load balance are generally can be elaborated as an optimization problem to maximize the utility function of network [7], [8]. In [7], the UA problem for load balance is formulated as an optimization problem of mixed integer nonlinear programming and a polynomial-time algorithm is designed to find a near-optimal solution. In [8], in consideration of the power control, load balancing, and user fairness, the nonconvex optimization problem about UA is presented which aims to maximize the network utility in cellular networks. Most of algorithms proposed by the previous studies are heuristic and adaptive under the unpredictable fluctuations of traffic. Hence, it is difficult to cope with the demand of surging traffic in UDN.

Traffic prediction methods can be divided into statistic-based methods and machine-learning-based methods. Most of the predicted models based on statistical methods use fitting models. For instance, autoregressive integrated moving average model (ARIMA) [9], generalized autoregressive conditional heteroscedasticity model (GARCH) [9] and Holt-Winter exponential smoothing [10] are used for traffic prediction. The fitting methods often employ self-similarity to forecast the traffic of cellular networks, so it is difficult to effectively forecast the large-scale nonlinear data of cellular networks. Aiming to capture the complex nonlinear characteristics of traffic data for cellular networks, the predicted models based on machine learning have increasingly become a powerful tool compared with classic statistical models.

It has been found that the mechanism of traffic prediction can be used to adjust the status of BSs [11]. A robust UA method is proposed in [12] under the intervention of the prediction mechanism and a pre-calculated robust framework is designed with unpredictable traffic spikes. In [13], using support vector regression (SVR) model, UEs can decide to associate with the appropriate cell under predicted demand

of traffic. The above methods either consider the throughput of system, or ensure that the BS is not overloaded, but it is not considered whether the overall load status of system is balanced or not in above studies. In [14], a hybrid deep learning model which includes autoencoder (AE) and long and short-term memory network (LSTM) is presented for traffic prediction. The management for UA is implemented based on the predicted results of load and traffic prediction has a great impact on the performance of wireless cellular networks [15].

B. Contributions and Organization

The contributions are the following:

- Firstly, taking into account the complex characteristics of wireless traffic in temporal and spatial dimension, a spatiotemporal feature based and densely connected convolutional neural network (STDCN) is proposed to forecast traffic fluctuations and the future UA status can be adjusted in advance according to the traffic forecasted.
- Then the problem for UA is elaborated as maximizing the utility function of load balancing index in order to realize the load balance of system in the long term and ensure quality of service (QoS) of UEs. The load-aware UA algorithm (i.e. DUALP) is proposed to solve the dual problem under the prediction mechanism.
- Finally, experimental results show that compared with the traditional method, DUALP realizes the overall load balance of the system and avoids the load imbalance of system as well as providing user's QoS guarantee and ensuring long-term stability of the system.

The rest of this paper is as follows. The system model and problem formulation are introduced in Section II. In Section III, traffic prediction based on machine learning is elaborated and the load-aware user association algorithm is proposed. Section IV gives the simulation results. Finally, the conclusion is summarized in Section V.

II. System Model And Problem Formulation

A. System Model

The area covered by the UDN is marked as $\mathcal{L}$. $\mathcal{B}$ is the set of small base stations (SBSs), where $\mathcal{B}$= $\{\mathrm{SBS}_1, \mathrm{SBS}_2, \cdots, \mathrm{SBS}_L\}$, and the number of SBS deployed is $L = |\mathcal{B}|$. At any location x in area $\mathcal{L}$, UEs generate data flow requests according to an inhomogeneous Poisson point process. The downlink of UDN is considered. The spatial intensity for data flow is $\lambda(x)$, and the mean of data packets is $\frac{1}{\mu}$. Therefore, $\frac{\lambda(x)}{\mu}$ captures the spatial traffic variability. It is assumed that UEs at the position x is associated with BS_i, the transmission rate provided by BS_i for UEs at the position x can be expressed as

$$C_i(x) = W \log\left(1 + SINR_i\left(x\right)\right), \tag{1}$$

where W is the available frequency bandwidth, the signal to interference plus noise ratio (SINR) can be expressed as

$$SINR_i(x) = \frac{P_i G_i(x)}{\sum\limits_{k \neq i} P_k G_k(x) + WN_0}, \forall i, k \in \mathcal{B}. \tag{2}$$

P_i represents the transmit power of BS_i, N_0 denotes the noise density, and $G_i(x)$ is the channel gain between UEs at location x and BS_i.

At location x, the association rule between UEs and BS_i is defined as $a_i(x)$. The constraint $\sum\limits_{i \in \mathcal{B}} a_i(x) = 1$ assumes that each UE can only be associated with one SBS. In order to ensure that all data flows generated by UEs at the position x can be served by BS_i, the resource block allocated by BS_i to UEs at the position x can be derived as $\frac{\lambda(x)}{uC_i(x)} a_i(x)$, thus the load of BS_i in the entire area $\mathcal{L}$ is

$$\rho_i = \int\limits_{\mathcal{L}} \frac{\lambda(x)}{uC_i(x)} a_i(x) dx. \tag{3}$$

The load represents the resource block required for BS_i to provide services for UEs associated with itself in the area $\mathcal{L}$.

B. Generic UA Problem: A Utility Maximization Perspective

It is assumed that the utility obtained is $U(\rho_i)$ when BS_i bears the load ρ_i where $U(\cdot)$ represents a class differentiable, monotonically increasing and strictly concave utility function.

In order to fully make use of the resource blocks which each BS provides, the UA problem can be described as maximizing the utility function of the load ρ_i. In UDN, an upper threshold is needed for each SBS with the consideration of dynamic fluctuation of traffic as well as UEs' QoS. Let c_i denote the maximum bearable load of BS_i, the constraint for UA is

$$\rho_i < c_i. \tag{4}$$

So, the UA problem can be formulated as

$$\begin{aligned} &\max_{a} \ U(\rho) \\ &\text{s.t.} \quad \begin{cases} \sum\limits_{i \in \mathcal{B}} a_i(x) = 1, \\ a_i(x) \in \{0, 1\}, \forall x \in \mathcal{L}, i \in \mathcal{B}, \\ 0 \leq \rho_i(a) < c_i. \end{cases} \end{aligned} \tag{5}$$

C. Logarithmic Utility Maximization for Load Balance

For the purpose of measuring the degree of load balance among different SBSs in the area $\mathcal{L}$, the Jain index to quantify the overall load status of the system in this area is introduced, that is

$$\eta = \frac{\left(\sum\limits_i \rho_i\right)^2}{L \sum\limits_i {\rho_i}^2}, \tag{6}$$

where L is the number of SBSs deployed, and ρ_i represents the load of BS_i. The logarithmic function is introduced as a utility function,

$$U(\eta) = \log(\eta) = 2\log(\sum_i \rho_i) - \log(\sum_i {\rho_i}^2) - \log(|\mathcal{B}|). \tag{7}$$

The last item in formula (7) does not affect the optimal solution of the problem. Therefore, the problem can be rewritten as

$$\max_{a} \left\{ 2\log\left(\sum_i \rho_i\right) - \log\left(\sum_i {\rho_i}^2\right) \right\}$$
$$\text{s.t.} \quad \begin{cases} \sum\limits_{i\in\mathcal{B}} a_i(x) = 1, \\ a_i(x) \in \{0,1\}, \forall x \in \mathcal{L}, i \in \mathcal{B}, \\ 0 \le \rho_i < c_i. \end{cases} \tag{8}$$

D. The Load Under Prediction Mechanism

Due to the error between the predicted value and the true value, so it is necessary to consider the impact of error on the final result of UA. An ideal model is chosen which regards the error $\varsigma_i(x)$ as a Gaussian random process.

Definition 1 : The predicted load of BS_i in the t -th slot is $\hat{\rho}_{i,t} = \int\limits_{\mathcal{L}} \frac{\hat{\lambda}_t(x)}{\mu C_i(x)} a_i(x) dx$, where $\hat{\lambda}_t(x)$ represents that BS_i predicts the spatial intensity of the data flow at location x in the t-th slot. For the prediction of $\hat{\lambda}(x)$, please see Section III, the subscript is omitted here.

Definition 2 : The predicted error $\varsigma_{i,t}(x)$, $\varsigma_{i,t}(x)$ is the error existing in the prediction process, and $\varsigma_{i,t}(x) \sim N(0, \sigma_1^2(x))$. $\sigma_1^2(x)$ is sample variance,

$$\sigma_1^2(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\rho}_{i,t}(x) - \bar{\rho}_i(x)\right)^2, \tag{9}$$

where $\hat{\rho}_{i,t}(x)$ is the predicted load in the t-th slot at location x, $\bar{\rho}_i(x) = \frac{1}{n} \sum\limits_{t=1}^{n} \hat{\rho}_{i,t}(x)$, $\bar{\rho}_i(x)$ represents the average of the predicted load under n time slots, n is total number of time slots. For ease of illustration, the subscript t is omitted later, so the actual load of BS_i in the future is

$$\tilde{\rho}_i(a_i(x)) = \hat{\rho}_i(a_i(x)) + \varsigma_i(x). \tag{10}$$

Relax $a_i(x) \in \{0,1\}$ to $a_i(x) \in [0,1]$, so the problem is rewritten as

$$\max_{a} \left\{ 2\log\left(\sum_i \tilde{\rho}_i\right) - \log\left(\sum_i {\tilde{\rho}_i}^2\right) \right\}$$
$$\text{s.t.} \quad \begin{cases} \sum\limits_{i\in\mathcal{B}} a_i(x) = 1, \\ a_i(x) \in [0,1], \forall x \in \mathcal{L}, i \in \mathcal{B}, \\ \tilde{\rho}_i(x) < c_i. \end{cases} \tag{11}$$

which is the final problem to be solved.

III. The UA Scheme Based on Traffic Prediction

A. Spatiotemporal Feature Based and Densely Connected Convolutional Neural Network

A new convolutional neural network (CNN) which is densely connected between convolutional layers based on the spatiotemporal feature is designed and called STDCN. STDCN includes the training data construction, CNN and parametric matrix based fusion.

Suppose the traffic data X_t in the t-th time slot is to be predicted, and the data before the t-th time slot is divided into two categories: the adjacent data and the periodical data. If the length of time before the t-th time slot is p, then the adjacent data is written as

$$X_p = [X_{t-p}, X_{t-(p-1)}, \cdots, X_{t-1}]. \tag{12}$$

Take the length of time in the previous q weeks, then the periodical data is written as

$$X_q = [X_{t-q*24}, X_{t-(q-1)*24}, \cdots, X_{t-24}]. \tag{13}$$

Then traffic data of two classes are concatenated to be a new tensor as the training dataset.

STDCN has two CNNs which have same architecture and parameters. Each CNN has k layers which implements a nonlinear transformation $f_k(\cdot)$ including convolution (Conv), batch normalization (BN) and the exponential linear unit (ELU). The feature of traffic in the spatial dimension is captured by the operation of convolution, and the feature in the time dimension is learned by two deep CNNs jointly. After k -layers feature learning, outputs of the two CNNs respectively are

$$\begin{aligned} X_p^k &= f_k\left(X_p^0 \oplus X_p^1 \oplus \cdots \oplus X_p^{k-1}\right) \\ X_q^k &= f_k\left(X_q^0 \oplus X_q^1 \oplus \cdots \oplus X_q^{k-1}\right) \end{aligned} \tag{14}$$

where $\oplus$ represents the concatenation operation of the feature maps produced in all preceding layers.

Since the predicted traffic data in different grids have different degrees of dependence on the adjacent data and the periodical data, weights of the adjacent data and the periodical data should also be different. Thus, a separate convolutional layer is added in $(k+1)$-th layer to get the features. The output after fusion for the feature of the adjacent data and the periodical data is

$$X_o = \omega_p \odot X_p^{k+1} + \omega_q \odot X_q^{k+1}, \tag{15}$$

where $\odot$ represents the Hadamard product, ω_p and ω_q represents the learnable parameter of the adjacent data and the periodical data, respectively. After the activation function, the final output of STDCN is

$$\hat{\lambda} = \sigma\left(X_o\right), \tag{16}$$

where σ represents activation function operation, $\hat{\lambda}$ is the predicted value obtained by STDCN.

B. Dynamic User Association Based on Load Prediction Algorithm

It can be seen that the problem described in formula (11) is an optimization problem with multiple equality constraints and inequality constraints which is difficult to solve directly. The idea of duality in convex optimization is used to solve the problem in this paper [16]. Because the Slater's Condition holds, thus satisfies the strong duality, that means the optimal solution of the dual problem is equal to the optimal solution of the primal problem.

Introducing the Lagrange multiplier β, the dual problem is

$$\min_{\beta \geq 0} \left\{ \max_{a} \varphi(a, \beta) \right\}$$
$$\text{s.t.} \quad \begin{cases} \sum\limits_{i \in \mathcal{B}} a_i(x) = 1, \\ a_i(x) \in [0,1],\ \forall x \in \mathcal{L}, i \in \mathcal{B}. \end{cases} \tag{17}$$

where the objective function is given by

$$\varphi(a, \beta) = 2 \log \left(\sum_i \tilde{\rho}_i \right) - \log \left(\sum_i \tilde{\rho}_i^{\,2} \right) + \sum_{i \in \mathcal{B}} \beta_i (\tilde{\rho}_i - c_i). \tag{18}$$

Define the feasible region of the optimization problem in formula (17) as

$$\mathcal{F} = \left\{ \begin{array}{l} \sum\limits_{i \in \mathcal{B}} a_i(x) = 1 \\ a_i(x) \in [0,1] \end{array} \right. , \forall x \in \mathcal{L}, \forall i \in \mathcal{B}. \tag{19}$$

The problem is composed of the inner problem and the external problem which is solved respectively. The method of Projected Gradient Descent (PGD) is used to solve the inner problem as shown in Algorithm 1 [17]. By implementing a PGD algorithm, the optimal UA rule a^* can be obtained for the inner problem. In order to solve the optimal solution of the inner and outer problem jointly, an averaging mechanism which averages the optimal solution generated by each PGD's iteration is proposed as shown in Algorithm 1. When the objective function $\varphi(a, \beta)$ converges, the optimal UA rule vector is obtained by Algorithm 1.

Algorithm 1 Dynamic User Association Based on Load Prediction Algorithm (DUALP)

Initialization: $a^{(0)}$, $\beta^{(0)}$, s
Iterate: over n, until convergence
  $\beta^{(n+1)} = [\beta^{(n)} + s(\widetilde{\rho} - c)]^+$
  PGD on $\mathcal{F}$:
  Iterate: over N, until convergence
    $t^{(N+1)} = a^{(N)} - s * \nabla_a \varphi(a^{(N)}, \beta^{(n+1)})$
    $a^{(N+1)} = \Pi_{splx}[t^{(N+1)}]$
    Where Π_{splx} represents the projection on $\mathcal{F}$:
    Sort $t^{(N+1)}$ in descending order:
    $t_1 > t_2 > \cdots > t_{|\mathcal{B}|}$
    Select $k = \arg\max\{j | t_j + \frac{1}{j}(1 - \sum_1^j t_i) > 0\}$
    $a_i^{N+1} = [t_i + \frac{1}{k}(1 - \sum_1^k t_i)]^+, i = 1, \cdots, |\mathcal{B}|$
  End PGD
${a_o}^{(n)} = \frac{1}{n} \sum_{i=0}^{n-1} a^{(i+1)}$
End

IV. Numerical Results

A. Preprocessing and Simulation Setup

The dataset of Telecom Italia is used for simulation in this paper [18]. The traffic data used in this paper is call detail records (CDRs) provided by Telecom Italia. The traffic data is recorded for two months during the period from 00:00 11/01/2013 to 00:00 01/01/2014. The time interval of dataset is 10 min, but a large part of the traffic in the grid is zero which will make the dataset very sparse, so the data in each grid is aggregated into 1 hour. The area covered by cell (5060) is the Duomo of Milan which is selected for traffic prediction. The Internet traffic activity are used to study for traffic prediction and UA. The Internet traffic activity of last seven days is divided into the testing set, and others are used as the training set in STDCN. The simulation parameters in this paper are shown in TABLE I where d represents the distance between BS and UEs.

TABLE I
Simulation Parameters

Parameters	Variable	Values
Covered Area (m^2)	$\mathcal{L}$	2350*2350
SBS Number	L	17
Transmission Power (W)	P	2
Path loss model	$PL(d)$	$PL(d) = \exp(-3d)$
Noise Density (dBm/Hz)	N_0	-174
Bandwidth (MHz)	W	10

B. The Performance of Different Prediction Mechanism

The results of different methods for traffic prediction within a week at Duomo are given. Three existing algorithms, ARIMA [9], Holt-Winter [10] and Densely Connected Convolutional Neural Network (DenseNet) [15], are adopted as baselines for comparison.

The predicted value and ground truth of different methods are shown in Fig. 1. Compared with the other three prediction methods, the result of STDCN matches the truth value of traffic trend very well even if the traffic data is unstable in the 48-th to 72-th slot. The method of ARIMA and Holt-Winter fail to stably capture the characteristics of dynamic changes for traffic data, especially the Holt-Winter model, because traditional statistical models usually use self-similarity to make prediction which is difficult to catch the temporal and spatial changes of traffic for UDN. Although DenseNet also has a great performance whereas its peak performance is weaker than STDCN.

The absolute cumulative error (ACE) is employed as the evaluation metric which is defined as

$$ACE(n) = \sum_{t=1}^{n} \left| \hat{\lambda}(t) - \lambda(t) \right|, \tag{20}$$

where $\hat{\lambda}(t)$ is the predicted traffic value, $\lambda(t)$ is the actual value, and n is the number of predicted data.

The result of ACE is shown in Fig. 2. Compared with the other predictive methods, the values of ACE obtained by STDCN are smaller in general. Hence, it can be seen that no matter what kind of the performance metric, STDCN has a relatively significant performance. This is because that the dependence of wireless traffic in the temporal dimension and spatial dimension have been jointly considered in STDCN proposed in this paper.

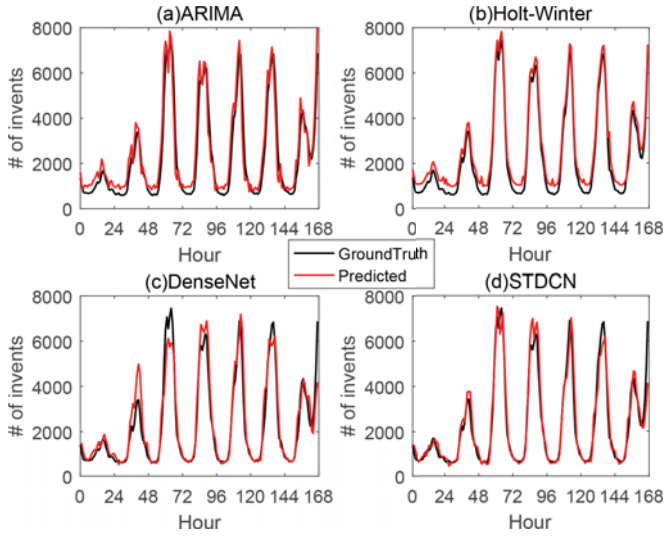


Fig. 1. The predicted value and ground truth.

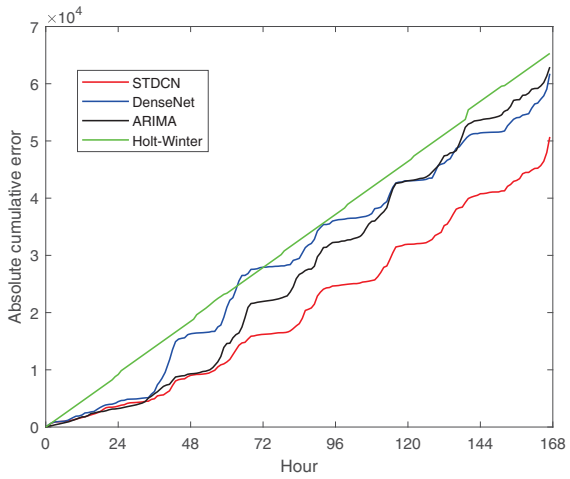


Fig. 2. Comparison of ACE.

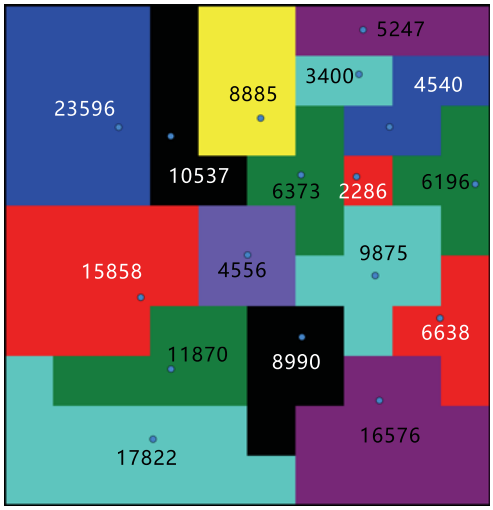
C. The User Association Map

A region covering 10*10=100 grids is selected to analyze the optimal UA map with the introduction of predictive mechanism and the result for performance is given. By applying machine learning and statistical tools on CDRs, we can determine the capacity requirements of UEs at different time periods.

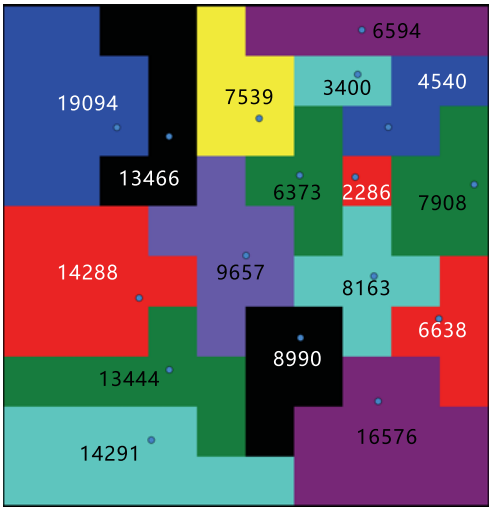
Given that the bias factor of BS_i is A_i, the bias SINR received by UEs at the position x is defined as

$$SINR_i(x)' = A_i \cdot SINR_i(x) = \frac{A_i P_i G_i(x)}{\sum\limits_{k \neq i} P_k G_k(x) + N_0}. \quad (21)$$

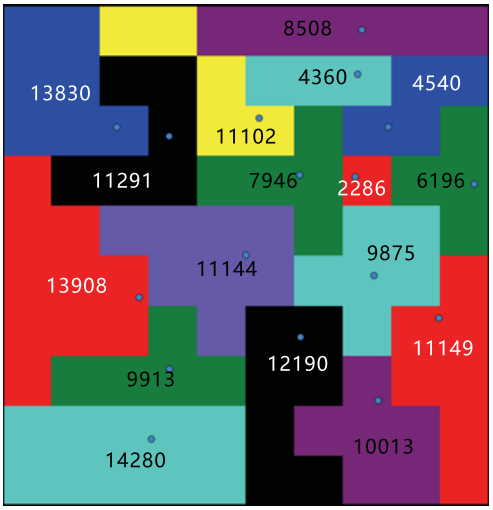
A method of SINR bias in which UEs select the SBS that can provide the highest bias SINR for association in comparison is considered. The bias value is set to greater than 1 for the lightly loaded SBS, so the coverage of the SBS will



(a) max-SINR



(b) SINR Bias



(c) DUALP

Fig. 3. The results of UA.

TABLE II
BIAS FACTOR

SBS_i	**Bias Value**
1	1.03
2,3,7,15,17	1.01
4,5,14	1.02
6,8,9,10,11,12,13,16	1.00

expand and the overloaded base station will release UEs which alleviates the pressure for those overloaded SBS. The bias factor of SBS was shown in Table II where the first column is the SBS' number and the second column is the value of bias factor.

Fig. 3(a) is the UA map achieved by max-SINR. Because all UEs in the UDN are served by SBSs and all SBSs has the same power settings, users will be associated with the nearest SBS. This will inevitably cause some SBSs to bear too much load while other SBSs are lightly loaded. Fig. 3(b) is the UA map based on SINR bias. Compared with the blue area in the upper left of Fig. 3(a), the coverage of this area has a significant decrease and the total traffic under this area has dropped by $\frac{23596-19094}{23596} = 19.09\%$ in Fig. 3(b) which indicates that the load of this SBS in Fig. 3(b) is allocated to the SBS of adjacent black area. The optimal solution can be obtained when the value of objective function in Algorithm 1 converges. The UA map which is optimal and more balanced in load distribution is shown in Fig. 3(c) by adjusting the UA rule. For example, compared to the UA scheme based on SINR bias, the coverage of SBS in the center area has been further expanded. It indicates that this SBS absorbs the load of the adjacent SBSs, especially the red area adjacent to the left and the green area below.

In table III, the results of three approaches for the average load balancing index in a week are showed. It indicates that compared with the max-SINR and SINR bias, DUALP can achieve load balancing more effectively, and it also proves that DUALP is a more effective load balancing algorithm in the medium and long-term scales.

TABLE III
THE AVERAGE OF LOAD BALANCING INDEX IN A WEEK.

Approaches	max-SINR	SINR Bias	DUALP
Load balancing index	0.774	0.814	0.912

V. CONCLUSION

Introducing the mechanism of traffic prediction into the study of the problem on UA will have a great impact on the system performance and the overall load status of the system. Traditional methods were incapable of capturing the fluctuation feature of traffic, so a new method which predicts the fluctuation of future traffic based on machine learning was proposed. Then the predicted traffic data obtained were used to dynamically adjust the status for UA. When the predicted traffic data had been obtained, DUALP was used to adjust the status of UA in advance which could achieve the optimal load distribution. Finally, the result showed that the most optimal load distribution is realized by DUALP which could guarantee the user's QoS on the medium and long-term scales. Besides, DUALP had significant load balancing capabilities which provides the long-term stability for UDN.

REFERENCES

[1] Q. Yan, W. Chen and H. V. Poor, "Big data driven wireless communications: A human-in-the-loop pushing technique for 5G systems," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 64-69, Feb. 2018.

[2] J. Liu, M. Sheng, L. Liu and J. Li, "Network densification in 5G: From the short-range communications perspective," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 96-102, Dec. 2017.

[3] B. Post and S. Borst, "Joint load-driven frequency allocation and user association in dense cellular networks," *2019 31st International Teletraffic Congress (ITC 31)*, Budapest, Hungary, 2019, pp. 75-83.

[4] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936-1947, Sept. 2017.

[5] K. Sun, J. Wu, W. Huang, H. Zhang, H. Hsieh and V. C. M. Leung, "Uplink performance improvement for downlink-uplink decoupled HetNets with non-uniform user distribution," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7518-7530, Jul. 2020.

[6] X. Ge, X. Li, H. Jin, J. Cheng and V. C. M. Leung, "Joint user association and user scheduling for load balancing in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3211-3225, May 2018.

[7] A. Alizadeh and M. Vu, "Load balancing user association in millimeter wave MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 2932-2945, June 2019.

[8] Sami, Mahmoud, and John N. Daigle, "User association and power control for UAV-enabled cellular networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 267-270, Mar. 2020.

[9] C. Ding, J. Duan, Y. Zhang, X. Wu and G. Yu, "Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility," *IEEE Trans. Intell. Transp. Systems*, vol. 19, no. 4, pp. 1054-1064, Apr. 2018.

[10] D. Tikunov and Toshikazu Nishimura, "Traffic prediction for mobile network using Holt-Winter's exponential smoothing," 2007 15th International Conference on Software, Telecommunications and Computer Networks, 2007, pp. 1-5.

[11] R. Li et al., "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175-183, Oct. 2017.

[12] N. Liakopoulos, G. Paschos and T. Spyropoulos, "Robust user association for ultra dense networks," *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, HI, 2018, pp. 2690-2698.

[13] Y. Qi and H. Wang, "QoS-aware cell association based on traffic prediction in heterogeneous cellular networks," *IET Commun.*, vol. 11, pp. 2775-2782, 2017.

[14] J. Wang et al., "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, Atlanta, GA, USA, 2017, pp. 1-9.

[15] C. Zhang, H. Zhang, D. Yuan and M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1656-1659, Aug. 2018.

[16] S. Boyd, and L. Vandenberghe, Convex Optimization. New York, NY, USA : Cambridge University Press, 2004.

[17] Y. Chen, and X. Ye, "Projection Onto A Simplex," Mathematics, 2011.

[18] Telecom Italia.Telecommunications - SMS, Call, Internet – MI. 2015. [Online]. Available: http://dx.doi.org/10.7910/DVN/EGZHFV