Machine Learning based Resource Allocation Strategy for Network Slicing in Vehicular Networks

Yaping Cui*, Xinyun Huang*, Dapeng Wu*, Hao Zheng*
*School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications, Chongqing, China
Email: cuiyp@cqupt.edu.cn, s190101129@stu.cqupt.edu.cn, wudp@cqupt.edu.cn, zhenghao0906@gmail.com

Abstract—To deal with the lack of prediction and management for vehicular network slice in existing research, this paper designs a machine learning based resource allocation strategy for vehicular network slicing. Firstly, a traffic prediction mechanism based on Convolutional Long Short-Term Memory (ConvLSTM) is proposed, which will capture the spatial-temporal dependencies of the traffic to predict traffic of complex slice services in the vehicular networks. Secondly, considering the imbalance of wireless resource utilization caused by the space-time difference between application scenarios, a shared proportional fairness scheme is proposed to achieve efficient and differentiated utilization of wireless resources. Finally, on the basis of ensuring the demand of each slice, the resource allocation algorithm based on the primal-dual interior-point method is used to solve the optimal slice weight allocation to minimize the system delay. Simulation results show that the service traffic prediction mechanism can be used to predict service traffic in the future. The average error rates of SMS, phone, and web traffic will be reduced, so that the user load distribution can be obtained a priori. Based on the predicted load distribution, slice weight distribution is performed in advance so that arranging delay is saved. The resource allocation algorithm based on the primal-dual interior-point method can well calculate the optimal slice weight distribution at this time.

Index Terms—Vehicular Networks, 5G Network Slicing, Traffic Prediction, Resource Allocation.

I. INTRODUCTION

Autonomous driving is one of the key scenarios in 5G networks. In order to achieve road safety of intelligent transportation systems (ITS), the ultra-reliable and low-latency communications (URLLC) must be guaranteed in vehicular networks. Moreover, for vehicle-to-everything (V2X) communications, the large data transmission of diversified service requirements poses the challenges to improve the transportation efficiency of ITS. Thus, the vehicular networks would be tailored to cater to these different requirements come from different services. Thus, multiple virtual networks are created via network slicing as a feasible way to meet it diverse needs [1].

Owing to the diverse needs of mobile users [2], academia and industry have reached a new consensus on the use of the machine (deep) learning and artificial intelligence (AI) [3] in 5G and evolved cellular networks. The International Telecommunication Union (ITU) has set up a special group to promote AI and machine learning to contribute to the efficiency of 5G systems [4]. The introduction of AI will enable wireless networks to self-optimize, which bring individual users and enterprises a more stable network connection that contributes to achieving better user experience and energy efficiency. In the process of slice resource allocation of realizing AI-enhanced vehicular communication networks, one of the basic problems lies in the accurate prediction of slice service traffic [5], and this is because many tasks in wireless communication require real-time or non-real-time traffic analysis. For example, the efficiency of demand-aware resource allocation tasks largely depends on the accurate prediction of future wireless traffic [6]. Also, the sleeping mechanism of functional base stations relies heavily on the predicted traffic of specific base stations or areas to achieve the purpose of green communication and satisfy the needs of users [7].

However, it is a very challenging task to predict the slice services simultaneously in the entire network at the same time. The specific reasons are as follows. First, mobile users have different needs of services at different times in different locations, which makes the service traffic difficult to predict. Secondly, user mobility makes spatial dependence exist in the slice service between geographically distributed cells [8], and this characteristic is particularly prominent in vehicular communication networks. Essentially, slice service prediction can be regarded as a time series predicting problem. According to the characteristics of the solution, the existing research can be roughly divided into two categories, i.e., statistics-based methods and machine learning-based methods. For the first category, statistical or probability distribution be used to model and predict slice traffic, including α-stable distribution, covariance function, and entropy theory [9]. However, some studies have pointed out that linear models are not suitable for many practical applications [10]. For the second category, with the accumulation of massive cellular traffic data and the development of machine learning and AI technology [11], the traffic prediction methods based on data-driven machine learning has become a strong competitor to traditional statistical models and get substantial attention in the field of communications [12]. In order to capture the spatial dependence of wireless traffic between geographically distributed base stations, literature [13] designed a hybrid deep learning model for spatiotemporal prediction, where the spatial dependence is modeled by autoencoder and the temporal dependence captured by Long Short Term Memory (LSTM). To simultane-
ously capture the spatial-temporal dependencies of traffic and predict the citywide traffic, literature [14] proposed a novel framework that uses a fusion strategy based on a parameter matrix to fuse different types of temporal dependence, and then introduce Convolutional Neural Networks (CNN) to learn spatial dependency and enhance feature propagation.

Based on the above research, this paper first uses ConvLSTM, which combines CNN and LSTM, to model the temporal and spatial dependencies of the slice service traffic in the vehicular communication network, and predicts various service traffic to obtain the results of user load distribution. Secondly, [15] designed a new wireless resource management scheme, i.e., Shared Proportion Fair Scheme (SPFS), to keep resource management in accordance with slicing user activity, and we use it for resource allocation. Finally, according to the system delay minimization problem under the proposed wireless resource management scheme, a slice resource allocation algorithm based on the primal-dual interior-point method is used to explore the optimal slice weight.

The system model is described as follows. Section II describes the system model and assumptions. Using the above model, a machine learning-based resource allocation strategy is presented in Section III, and the resource allocation problem is formulated as a convex optimization problem to obtain optimal slice weight allocation. In section IV, we propose a resource allocation algorithm based on the primal-dual interior-point method to solve the optimal slice weight problem while minimizing the system delay. Simulation results are provided in Section V, followed by concluding remarks in Section VI.

II. SYSTEM MODEL

We consider the vehicular networks consist of $B$ Road Side Units (RSU) and $V$ network slices. And the sets of RSUs and slices are denoted by $B$ and $V$, respectively. As shown in Fig.1, the function of RSU is virtualized into three layers, namely RSU interface layer, RSU virtualization layer and RSU virtual resource layer. Among them, the RSU interface layer provides related interfaces for each slice. The RSU virtualization layer implements the functions of slice management, SDN control, and slice coordination. Besides, the RSU virtualization layer provides the virtual resources required by each slice, which are obtained from the resource sharing layer of the base station. Also, the base station controls the service traffic prediction of the entire system and schedules the resources among slices.

System states $U_b^c$, $U_b^v$, and $U^v$ represent the sets of vehicles that communicate with RSU $b$ on slice $v$, communicate with RSU $b$, and on slice $v$, respectively. Specially, we use $n_b^c$ and $n^v$ denote the cardinalities of these sets, i.e., $|U_b^c| = n_b^c$, and $|U^v| = n^v$. Further, we assume that each vehicle communicates with only one RSU and connects to one slice.

In our model, the vehicle communicates with the RSU that provides it with the strongest SINR, and the downlink SINR can be expressed as

$$\text{SINR}_{ub} = \frac{P_b G_{ub}}{\sum_{k \in B} P_k G_{uk} + \sigma^2}$$  \hspace{1cm} (1)

where the spectrum of the RSU is set to 20 MHz, and the transmit power $P_b$ is set to 44 dBm [16]. The channel gain $G_{ub}$ is related to path loss, shadow fading, and fast fading. The path loss is equals to $39 \log_{10}(d_{ub}) + 25 + 20 \log_{10}(f_c)$ [17], where $d_{ub}$ and $f_c$ represent the distance between the vehicle and the RSU and carrier frequency, respectively, and we set $f_c = 4$ GHz. Shadow fading follows logarithmic normal distribution with mean square deviation 4 dB. Fast fading is Rayleigh distribution depending on vehicles speed. The noise $\sigma^2$ depends on the noise spectral density $\eta = -174$ dBm/Hz and the noise figure $\gamma = 9$ dB.

According to the Shannon capacity formula, we get the spectrum efficiency of vehicle $u$ at RSU $b$, which can be expressed as

$$e_{ub} = \log_2(1 + \text{SINR}_{ub})$$ \hspace{1cm} (2)

The limited resources of each RSU are shared by all connected vehicles, so vehicle $u \in U_b$ can be allocated a fraction of resources from RSU $b$. The allocated resources may be some resource blocks or time slots. For simplicity, we use $f_u \in [0, 1]$ to represent the allocated resource for vehicle $u$, which implies a proportion of total resource of RSU so that $\sum_{u \in U_b} f_u = 1$. Therefore, the transmission rate of RSU $b$ to vehicle $u$ is formulated as $r_u = f_u c_u$, where $c_u$ denotes transmission rate when all resources of RSU $b$ are allocated to vehicle $v$, and $B$ is bandwidth.

III. WIRELESS RESOURCE ALLOCATION STRATEGY

In this section, we will describe the machine learning based resource allocation strategy. The cellular traffic prediction architecture based on ConvLSTM is firstly introduced. Then, we will develop the resource allocation scheme for the radio resource allocation issue. In addition, the system delay minimization problem will be formulated at the end of this section.

A. Cellular Traffic Prediction Architecture

ConvLSTM neural networks are adapted to predict service traffic. ConvLSTM networks can not only model the sequences information of cellular traffic accurately as same as Long Short-Term Memory networks (LSTM), but also the local feature as same as convolutional neural networks (CNN). In short, it can easily capture the spatial-temporal dependencies. Consequently, ConvLSTM networks are appropriate for predicting slice traffic in complex vehicular networks. This memory cell consists of cells states and three neural network units, i.e., input gate, forget gate, as well as output gate. For this specific framework, it is able to effectively store information chronically from long-term sequences.

As shown in Fig. 2, the forget gate outputs a value $f_t \in [0, 1]$ to the cell under the current input $x$ and past cell output $H_{t-1}$, which determines what information should be abandon in past cell status $C_{t-1}$ now. The calculation formula of forget gate is given by

$$f_t = \sigma(W_f^x x_t + W_f^h H_{t-1} + b_f)$$ \hspace{1cm} (3)
The input gate decides update when a new input comes to the ConvLSTM unit through a sigmoid function, which can be further effects present states $C_t$, and expressed as

$$i_t = \sigma(W^i_x \ast x_t + W^i_h H_{t-1} + b_i)$$ (4)

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W^c_x \ast x_t + W^c_h H_{t-1} + b_c)$$ (5)

The output gate decides the output of this cell through a sigmoid function. After that, the output $H_t$ is given by

$$o_t = \sigma(W^o_x \ast x_t + W^o_h H_{t-1} + b_o)$$ (6)

$$H_t = o_t \odot \tanh(C_t)$$ (7)

where $f_t$, $i_t$, $o_t$, $C_t$, $H_t$, denote the output of forget gate, the output of input gate, cells status, cells input and cells output, respectively. Different from the common LSTM networks, the inputs or outputs in the ConvLSTM unit are all three-dimensional tensors. More specifically, the citywide service traffic data can be deemed treated as a matrix or picture. Then, previous multiple service data are fed into the ConvLSTM networks to obtain future results. So, the multiply operation of common LSTM networks is replaced by convolution operation in ConvLSTM. Thus, in the above equations, notation $\ast$ denotes the convolution operation and notation $\odot$ denotes the Hadamard product. The neural network can be accomplished via updating various parameters in each iteration, e.g., $W^x_t$, $W^f_t$, $W^c_t$, $W^o_t$, $W^h_t$, $W^c_h$, $W^o_h$, $W^h_h$, so that the networks can minimize the error between forecasted values and ground trues.

### B. Shared Proportional Fairness Scheme

For each slice in vehicular networks, we assume each slice is allocated a certain percentage of the radio resources, which is denoted by $s_v^u, v \in \mathcal{V}$, so we have $s_v^u > 0, \forall v \in \mathcal{V}$, and $\sum_{v \in \mathcal{V}} s_v^u = 1$. Next, the vehicle gains sub-weight from the serving slice depends on the number of active vehicles, i.e., for a vehicle $u \in \mathcal{U}^v, \forall v \in \mathcal{V}$, $w_u = \frac{s_v^u}{n_v}$. where $w_u$ means the sub-weight of vehicle $u$. Finally, the RSU allocates its resources to vehicles in proportion to their weights. Consequently, the transmission rate from RSU $b$ to vehicle $u$ can be obtained and written as

$$r_u = \frac{w_u b}{\sum_{u' \in \mathcal{U}} w_{u'}} c_u = \frac{n_v s_v^u}{n_v^2} c_u$$ (8)

Considering there are many vehicles on slice $v$ at RSU $b$, so the average transmission rate based on some further notations that are introduced in Table I is expressed as

$$\bar{r}_b^v = \frac{s_v^u}{g_b} c_u$$ (9)

### TABLE I: KEY NOTATIONS

| Notation | Definition | Interpretation |
|----------|------------|----------------|
| $\rho^o$ | overall load of slice $v$ | Load distribution of slice $v$ |
| $\rho^r$ | load distribution of slice $v$ | Relative load distribution of slice $v$ |
| $\tilde{g}$ | overall weight relative load distribution | Overall weight relative load distribution |
| $\delta^v$ | diagonal matrix of mean reciprocal capacity of slice $v$ | Mean reciprocal capacity of slice $v$ |
| $\Delta^\rho$ | diagonal matrix of mean reciprocal parameter of slice $v$ | Mean reciprocal parameter of slice $v$ |

According to (9), so the average bit transmission delay (BTD) of the vehicle on slice $v$ can be given by

$$\text{BTD}^v = \sum_{b \in \mathcal{B}} \tilde{p}_b^v \text{BTD}_b^v = \frac{\rho^v (\tilde{\rho}^v, \tilde{g}) \Delta^\rho}{s^v}$$ (10)
where BTD\textsuperscript{v}_b represents the average BTD of the vehicle at RSB b on slice v. Besides, we use \( \langle x_1, x_2 \rangle_M \triangleq x_1^T M x_2 \) and \( \| x \|_M \triangleq \sqrt{x^T M x} \) denote the weighted inner product of the vectors and the weighted norm of a vector. M denotes a diagonal matrix.

We assume the task process follows the GI/M/1/∞ queue model [18]. Among which, the random variable of task arrival interval obeys general distribution \( F(t) \), \( t \geq 0 \), while \( F(t) \) in the different time slots is independent and identically distributed. Its expectation \( \frac{1}{\lambda} = \int_0^\infty t dF(t) \), \( \lambda > 0 \), where \( \lambda \) is arrival rate. The task service time is exponentially distributed, i.e., \( G(t) = 1 - e^{-\lambda t}, \ t \geq 0 \), and the mean value of the task service times depends on the allocated resource blocks for the vehicle from RSU. To facilitate the analysis of the system delay, \( \frac{1}{\mu} \) is used to denote the average service time when only one resource block is available for the task. Similarly, we denote the average service time as \( \frac{1}{\text{RB}_b} \) when RB\textsuperscript{v}_b resource blocks for the task. The average waiting delay of these vehicles \( u^v_b, u^v_i \in \mathcal{U}^v_b \) is given by

\[
WD^v_b = \frac{1}{\text{RB}_b^v} (1 - \sigma_b^v) \tag{11}
\]

where \( WD = \sum_{v \in \mathcal{V}} u^v_b e^{-\lambda t} \), and \( \sigma_b^v \) can be obtained by solving the following equation:

\[
\int_0^\infty e^{-RB^v_b} (1-\sigma_b^v t) \frac{\rho^v_b e^{-\rho^v_b t} + \rho^v_b e^{-\lambda t} e^{-\rho^v_b t} - \rho^v_b e^{-\lambda t}}{1 - e^{-\rho^v_b}} dt = \sigma_b^v \tag{12}
\]

As a result, the average waiting delay of the vehicle on slice \( v \) is

\[
WD^v = \sum_{b \in \mathcal{B}} \sigma_b^v WD^v_b = \frac{\rho^v \langle \tilde{\rho}^v, \tilde{g} \rangle \theta_u}{s^v \mu^v} \tag{13}
\]

According to formula \( \langle x_1, x_2 \rangle_{M_1 + M_2} \triangleq \langle x_1, x_2 \rangle_{M_1} + \langle x_1, x_2 \rangle_{M_2} \) and (10) (13), the total average delay of vehicles on slice \( v \) can be obtained by

\[
D^v_{\text{Total}} = BTD^v + WD^v = \frac{\rho^v \langle \tilde{\rho}^v, \tilde{g} \rangle \mu^v \Delta_+ + \theta^v}{s^v \mu^v} \tag{14}
\]

\[C. \ Problem \ Formulation\]

In the real implementation of network slicing, each slice would provide a guarantee of service for the vehicles, that is, the total delay on slice \( v \) does not exceed a deadline \( d_v \). In this subsection, we will explore how to obtain the optimal solution of minimizing system delay by allocating weight to each slice.

Considering a network with just one slice \( v \), so we have \( s^v = 1 \), \( \tilde{g} = \tilde{\rho}^v \). To satisfy the deadline \( d_v \), from (14), we can obtain that

\[
\rho^v \leq \ell (d_v, \tilde{\rho}^v) \triangleq \frac{\mu^v d_v}{\| \tilde{\rho}^v \| \mu^v \Delta_+ + \theta^v} \tag{15}
\]

where we define \( \ell (d_v, \tilde{\rho}^v) \) as the acceptable maximum load of slice \( v \).

Next, considering a multi-slice networks, and each slice has itself requirements. According to (14) (15), each slice should satisfy the following constraint to meet these requirements: \( \forall v \in \mathcal{V} \)

\[
s^v \geq \ell (d_v, \tilde{\rho}^v) - \rho^v \sum_{u \neq v} s^u \langle \tilde{\rho}^u, \tilde{\rho}^v \rangle (\mu^v \Delta_+ + \theta^v) \tag{16}
\]

Equation (16) can be written in a simplified form, i.e.,

\[
\sum_{v \in \mathcal{V}} s^v h^v \geq 0 \tag{17}
\]

where \( h^v = (h_u^v : u \in \mathcal{V}) \) is share coupling vector of slice \( v \) and can be expressed by

\[
h^v = \begin{cases}
1 & v = u \\
\frac{\rho^v}{\| \tilde{\rho}^v \| \mu^v \Delta_+ + \theta^v} & v \neq u
\end{cases}
\]

Our objective is to satisfy the requirements of vehicles in each slice and minimize the system overall delay. Consequently, the optimization problem can be formulated as

\[
\min \quad \frac{\rho^v \langle \tilde{\rho}^v, \tilde{g} \rangle \mu^v \Delta_+ + \theta^v}{s^v \mu^v} \\
\text{s.t.} \quad C1 : \sum_{v \in \mathcal{V}} s^v h^v_i \geq 0 \quad i = 1, 2, \cdots, V \tag{18}
\]

\[
C2 : \sum_{i=1}^V s^i = 1
\]

\[
C3 : s^i \geq 0 \quad i = 1, 2, \cdots, V
\]

where \( s^v \) is the optimization variable. Constraint C1 ensures the requirements of slices can be satisfied. Constraints C2 and C3 state the weight of each slice is nonnegative and is constrained by total resources. In practice, looking for the optimal weights of each slice through minimizing the system total delay according to the current load distribution will cause some delay, we define that as arranging delay DARR. Through proposed traffic prediction architecture, the system can acquire complicated traffic information in advance to calculate the optimal weights for each slice. Hence, the arranging delay can be largely reduced.

IV. PRIMAL-DUAL INTERIOR-POINT METHOD BASED RESOURCE ALLOCATION ALGORITHM

The algorithm includes two phases, the first phase is service traffic prediction by using machine learning described in preceding section, and the second phase is the optimizing procedure which based on the primal-dual interior-point method.

Considering there are inequality constraints in convex optimization problem (18), so it can be solved by the primal-dual interior-point method. We will transform the inequality constrained problem as an equality constraint problem so that the central path of this problem can be found. Therefore, we
rewrite the problem (18) as
\[
\min \sum_{v \in V} \frac{\rho_v (\phi_v - \hat{\phi}_v) \mu_v \Delta_v + \theta_v}{s_v^T \mu_v} + \sum_{i=1}^V \left( -\log \left( \sum_{v \in V} s_v^T h_i \right) \right) t_i \\
+ \sum_{i=1}^V -\log (s_i^T) t_i
\]
\[
s.t. \sum_{i=1}^V s_i = 1, \quad i = 1, 2, \cdots, V
\]

For simplicity, let \( f_0(x) = \sum_{v \in V} \frac{\rho_v (\phi_v - \hat{\phi}_v) \mu_v \Delta_v + \theta_v}{s_v^T \mu_v}, \phi(x) = \sum_{i=1}^V \left( -\log \left( \sum_{v \in V} s_v^T h_i \right) \right) + \sum_{i=1}^V -\log (s_i^T) \), where \( x = [s_1, s_2, \cdots, s_V, -f_0(x)] \) is the optimization variable. Considering the equivalent problem
\[
\min f_0(x) + \phi(x)
\]
s.t. \( Ax = 1 \) (20)

**Algorithm 1** Resource allocation algorithm based on the primal-dual interior-point method

**Input:** \( J \) times observed load distribution \( \rho_{t-J}, \rho_{t-J+2}, \cdots, \rho_t \), initial \( x_0 \), \( \lambda_0 \), scale factor \( k \), residual error \( \epsilon_{feas} \), duality gap error \( \epsilon \)

**Output:** Optimal solution \( x^*_t+1, \cdots, x^*_{t+K} \)

**Phase 1:** Predict service traffic
Training the ConvLSTM networks to obtain parameters \( W^f, W^x, W^h, b_f, b_g, b_o \)

According to \( j \) times observed load distribution predict \( K \) sequences in the future \( \hat{\rho}_{t+1}, \cdots, \hat{\rho}_{t+K} = \arg \max p(\rho_{t+1}, \cdots, \rho_{t+K}) \)

**Phase 2:** To obtain optimal slice weight
while True do

\[
\text{Calculate initial value of the surrogate gap } \eta \leftarrow f(x)^T \lambda \text{ if } (\|\gamma_{pri}\| < \epsilon_{feas}) \& \& (\|\gamma_{dual}\| < \epsilon_{feas}) \& \& (\eta < \epsilon) \text{ then break end if}
\]

Determine \( t \leftarrow 2kV/\eta \)
Compute primal-dual search direction \( \Delta y_{pri} \)
Determine initial step length \( s_0 = \min \{0.99, \min \{-\lambda_i/\Delta \lambda_i | \Delta \lambda_i < 0\}\} \)
while \( \min \{f_i(x+s\Delta x) | i = 1, \cdots, 2V\} \) do

Ensure satisfy the constraint condition \( s \leftarrow \beta s \)
end while

while \( \|\gamma_{pri}(x+s\Delta x, \lambda + s\Delta \lambda, v + s\Delta v)\| = (1 - \alpha s) \|\gamma_{pri}(x, \lambda, v)\| \) do

Determine backtracking search step length \( s \leftarrow \beta s \)
end while
Update search direction \( y \leftarrow y + \Delta y_{pri} \)
end while

Due to the number of slices is three in simulation setting, so \( x = [s^1, s^2, s^3, -f_0(x)] \) and \( A = [1, 1, 1, 0] \). According to reference [19], the Newton step \( \Delta y \) is given by the modified KKT equations
\[
\begin{bmatrix}
\nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla^2 f_i(x) & D f(x)^T & A^T \\
-\text{diag}(\lambda) & -\text{diag}(f(x)) & 0 \\
A & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\Delta \lambda \\
\Delta v \\
\Delta y
\end{bmatrix}
= \begin{bmatrix}
\gamma_{dual} \\
\gamma_{pri}
\end{bmatrix}
\]

where \( \lambda_i = -\frac{1}{tf_i(x), i = 1, \cdots, 2V, f(x) = \begin{bmatrix} f_1(x) \\
\vdots \\
f_{2V}(x) \end{bmatrix} \)

are the inequality constrained functions of the original problem, \( D f(x) = \begin{bmatrix} \nabla f_1(x)^T \\
\vdots \\
\nabla f_{2V}(x)^T \end{bmatrix} \) is its derivative matrix. \( \gamma_{dual} \) and \( \gamma_{pri} \) are dual residual and primal residual, respectively, and these residuals are used for the termination condition in the primal-dual interior-point method. The solution of (21) is the current primal-dual search direction \( \Delta y_{pri} = (\Delta y_{pri}, \Delta \lambda_{pri}, \Delta v_{pri}) \).

The step length can be obtained by backtracking line search which based on the norm of the residual. Through continuous iteration, the best solution will be returned when the present state satisfies the termination condition, the details are elaborated in Algorithm 1.

V. PERFORMANCE EVALUATION

In this section, we use the cellular traffic dataset [20] of Milan, Italy to train our neural networks, this dataset contains three categories service traffic, i.e., SMS, phone, web traffic, and we can deem it as three slices. The area of Milan is divided into a grid overlay of 100 x 100 squares, and the dataset records 1000 samples of each square with an temporal interval of 1 hour. We choose the 800 samples from this dataset as the training set, the rest of the samples as the testing set. After 100 times iteration, the trained model is used for predicting the three types of service traffic. Then, the resource allocation algorithm based on the primal-dual interior-point method is used to solve the optimal slice weight allocation according to the forecast. In simulation, we set scale factor \( k = 2 \), residual error \( \epsilon_{feas} = 10^{-6} \), duality gap error \( \epsilon = 10^{-8} \).

Fig. 3 shows the network slicing delay with and without prediction when slice weight is fixed, and the default of arranging delay \( D_{ARR} \) was set to 20 ms. It can be seen that the periodicity of service traffic dynamically mapped to system delay completely. Using the service traffic prediction mechanism, the slice weight can be distributed on average 15.33 ms in advance.

The predicted results express system load distribution in the future, the optimal slice weight can be solved accordingly. In Fig. 4, we present the optimal slice weights under various times. Specially, at the 12th, 36th, 60th, and 84th time slots, the optimal slice weights of SMS are 0.3011, 0.3028, 0.3001, and 0.3076, respectively. Similarly, at these time slots, the
has a perfect effect on slice resource allocation.

**REFERENCES**

[1] S. Abdelwahab, B. Hamdaoui, M. Guizani and T. Znati, “Network function virtualization in 5G,” in *IEEE Communications Magazine*, vol. 54, no. 4, pp. 84-91, Apr. 2016.

[2] F. Xu, Y. Li, H. Wang, P. Zhang and D. Jin, “Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment,” in *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147-1161, Apr. 2017.

[3] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255-260, Jul. 2015.

[4] ITU News. “ITU launches new Focus Group to study machine learning in 5G systems,” [online]. Available: https://news.itu.int/itu-launches-new-focus-group-study-machine-learning-5g-systems, Accessed on: June. 29, 2018.

[5] P. K. Agyapong, M. Iwamura, D. Stuehle, W. Kiess, and A. Benjebbour, “Design considerations for a 5G network architecture,” *IEEE Communications Magazine*, vol. 52, no. 11, pp. 65-75, Nov. 2014.

[6] C. Yao, C. Yang, and C.-L. I, “Data-driven resource allocation with traffic load prediction,” *Journal of Communications and Information Networks*, vol. 2, no. 1, pp. 52-65, Mar. 2017.

[7] J. Wu, Y. Zhang, M. Zukerman, and E. K. N. Yang, “Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey,” *IEEE Communications Surveys and Tutorials*, vol. 17, no. 2, pp. 803-826, Second Quarter 2015.

[8] H. Wang, J. Ding, Y. Li, P. Hui, J. Yuan, and D. Jin, “Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks,” in *Proceedings of the 7th International Workshop on Hot Topics in Planet-scale mObile Computing and Online Social networking*, ser. HOTPOST 15. New York, NY, USA: ACM, Jun. 2015, pp. 19-24.

[9] R. Li, Z. Zhao, X. Zhou, J. Palicot, and H. Zhang, “The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice,” *IEEE Communications Magazine*, vol. 52, no. 6, pp. 234-240, Jun. 2014.

[10] J. G. D. Gooijer and R. J. Hyndman, “25 years of time series forecasting,” *International Journal of Forecasting*, vol. 22, no. 3, pp. 443-473, Apr. 2006.

[11] J. Wang, B. Gong, H. Liu, and S. Li, “Multidisciplinary approaches to artificial swarm intelligence for heterogeneous computing and cloud scheduling,” *Applied Intelligence*, vol. 43, no. 3, pp. 662-675, Oct. 2015.

[12] L. Nie, D. Jiang, S. Yu, and H. Song, “Network traffic prediction based on deep belief network in wireless mesh backbone networks,” in *IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1-5.

[13] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, “Spatiотemporal modeling and prediction in cellular networks: A big data enabled deep learning approach,” in *IEEE Conference on Computer Communications (INFOCOM)*, Atlanta, GA, USA, May 2017, pp. 1-9.

[14] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, “Citywide cellular traffic prediction based on densely connected convolutional neural networks,” *IEEE Communications Letters*, vol. 22, no. 8, pp. 1656-1659, Aug. 2018.

[15] J. Zhang, P. Caballero, G. de Veciana, S. J. Baek and A. Banchs, “Statistical Multiplexing and Traffic Shaping Games for Network Slicing,” in *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2528-2541, Dec. 2018.

[16] 3GPP. TR 38.802 v14.2.0. Technical Specification Group Radio Access Network; Study on new radio access technology physical layer aspects. 2017.

[17] Y. J. Bulitide, T.Rautiainen, “IST-4-027756 WINNER II D1. 1.2 V1, 2 WINNER II Channel Models,” EBITG, TUL, OUULU, CU/CRC, NOKIA, Tech. Rep. 2007.

[18] L. Kleinrock, Queueing Systems. Volume I: Theory. New York, NY, USA: Wiley, 1972.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[20] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, “A multi-source dataset of urban life in the city of Milan and the Province of Trentino,” *Scientific Data*, vol. 2, p. 150055, Oct. 2015.