电子科技大学
格拉斯哥学院
Glasgow College, UESTC

Final Year Project Report

Bachelor of Engineering

Optimal spectrum utilisation based on energy efficient cell switching in 5G ultra dense networks

Student: Lianrui Yang

GUID: 2357870

1st Supervisor:

2nd Supervisor:

2020-21

Coursework Declaration and Feedback Form

The Student should complete and sign this part

Student Name: Lianrui Yang	Student GUID: 2357870
Course Code : UESTC4006P	Course Name : INDIVIDUAL PROJECT 4
Name of 1st Supervisor:	Name of 2nd Supervisor:
Title of Project: Optimal spectrum utilisation based on energy efficient cell switching in 5G ultra dense networks	
Declaration of Originality and Submission Information	
I affirm that this submission is all my own work in accordance with the University of Glasgow Regulations and the School of Engineering requirements Signed (Student) : Lianrui Yang 杨连瑞	UESTC4006P

I affirm this submission is completed by the student independently and the quality of the submission meets the requirements for graduation. I consent to the student taking part in the oral presentation. Signed (1st Supervisor) :	

<table>
<tr><td colspan="2">Date of Submission : 23 April 2021</td></tr>
<tr><td colspan="2">Feedback from Lecturer to Student – to be completed by Lecturer or Demonstrator</td></tr>
<tr><td colspan="2">Grade Awarded:
Feedback (as appropriate to the coursework which was assessed):</td></tr>
<tr><td>Lecturer/Demonstrator:</td><td>Date returned to the Teaching Office:</td></tr>
</table>

Abstract

The fifth generation of mobile networks (5G) utilizes Ultra-dense network (UDN) that contains many base stations (BSs), which consumes a large amount of energy. One solution to save energy is idling BSs with low traffic demand and off-load these data traffic to Macro BS. Once a BS owned by a primary mobile network operator (MNO) is idled, its spectrum can be leased to a secondary MNO for earning revenue. This paper concentrates on applying reinforcement learning (RL) to maximize energy-saving and spectrum leasing revenue. Furthermore, a clustering-based method is proposed to mitigate the drawbacks of the RL algorithm. Results show that for a 5G system which contains 13 BSs, RL can reach 50% of the highest revenue and reduce calculating demand significantly compared with the traditional optimizing method. Clustering-based method can be applied on a system containing 101 BSs, which cannot be solved by RL. 0.46% of total energy consumption is saved by clustering.

Keywords: 5G, Ultra-dense network, Spectrum leasing, Energy Saving, Reinforcement learning, Multi-armed bandit problem, Clustering

Acknowledgements

I am very grateful to my professor and his doctor student for their support during the stressful time while I was designing the project and writing this report.

Contents

1 Introduction

1.1 The fifth generation of mobile networks and Ultra-dense network

The fifth generation of mobile networks is called 5G. The data transmitting rate of the 5G system is roughly and theoretically 1000 times larger than the 4G system [1]. To tackle such a large amount of data rate demand and enhance spectrum efficiency, the 5G system adopts the ultra-dense network (UDN). Ultra-dense heterogeneous networks (UDHetNet) contain a significant number of different kinds of small cells (SCs) such as microcell, picocell, and femtocell [2] to meet ever-growing communication requirements in different geographical areas.

1.2 Energy saving requirement and control-data separation architecture (CDSA)

Although the energy efficiency of 5G is higher than 4G, 5G's overall energy consumption demand is increasing. It causes carbon emission [3], [4] because the number of base stations in the 5G system is significantly larger than the 4G system [5]. BSs have become the primary power-consuming devices in the mobile network. Meanwhile, more antennas, millimetre wave (mmWave) signal [6], and larger transmitting bandwidth [7] deter this problem. In recent years, an environmentally friendly communication system has become a hotpot in mobile communication designing.

It is reported that almost 2% of greenhouse gas emit is caused by mobile communication [8] and radio access networks consume at least 60%-80% of mobile communication energy [9], [10]. Although there are many other methods to increase power efficiency, the most common approach to save energy is dynamically switching off small BSs that have low traffic demands and off-load these demands to Macro BSs. The control-data separation architecture (CDSA) of the mobile wireless access network allows switching on/off small BSs randomly. [11] introduces CDSA in detail. CDSA has a control base station (CBS) which is the Macro BS in 5G system. Other small BSs are called database stations (DBSs). All DBSs are connected with CBS by wire. A reliable algorithm can be executed on CBS to control DBSs in active or idle mode dynamically. The user equipment can only communicate with CBS and those DBSs with active mode. To reduce energy consumption, designing a perfect algorithm is important.

1.3 Spectrum leasing requirement

As a kind of natural resource, spectrum is valuable. On the one hand, 5G requires more spectral resources such as millimetre wave (mmWave) to reach a higher data rate. On the other hand, valuableness and scarcity of spectrum limit enlarging bandwidth. Those mobile network operators (MNOs) who possess spectrum are called primary mobile network operators (Primary MNOs). Other smaller MNOs who cannot afford to own spectrum and have to lease spectrum from primary MNOs are secondary mobile network operators (Secondary MNOs). As [12] reported, the spectrum owned by primary MNOs is not always adequately utilized. One solution to this waste is that primary MNOs can lease spectrum to secondary MNOs while their traffic demand is low. Therefore, leasing spectrum to secondary MNOs can not only increases the income of primary MNOs but also reach the utmost of utilizing spectrum. CDSA also allows spectrum leasing.

1.4 Possible solutions of energy saving and spectrum leasing

Generally, there are two main directions of designing a control algorithm of CDSA. They are traditional resource management algorithm and machine learning. The limitations of the traditional resource management algorithm are twofold. Firstly, it cannot optimize the problem globally. Secondly, because of a large amount of calculation, the result may not be obtained in real-time [13]. These two drawbacks limit its application. Oppositely, machine learning can avoid these problems. However, machine learning cannot always find the best solution.

2 Related works

2.1 Traditional methods of energy saving methods

To mitigate drawbacks of cell switching and help 5G system becomes more sustainable, in [14], Alsafasfeh. Q et al. designs a basic small-cell BS switching power control system to switch BSs in terms of time. After processed by their algorithm, the number of active BSs decreases, and power efficiency is higher than the conventional coordinated beamforming approach. In [15], the authors provide switching on/off algorithm based on Interference Contribution Rate (ICR). Signal interferences are calculated dynamically to determine a set of small BSs that should be switched off. Conclusions show that the algorithm can both increases energy efficiency and alleviates interferences between small cells. Y. Li et al. [16] provide a probability-based method on BS switching. They apply Poisson Point Process (PPP) distribution to model data transmission and Signal to Interference plus Noise Ratio (SINR) of Macro BS and small BSs. Average power consumption is calculated by the data transmission model. Results show that at least 30% of total consumed energy can be saved by this algorithm.

2.2 Reinforcement learning based energy saving methods

Owning to the development of computational technology, it is possible to optimize BS switching problem by machine learning (ML). [5] provides a series of mathematical system model including energy consumption model, system access model, etc. and applies a kind of ML cannel genetic algorithm to provide a dynamically sleeping BSs solution.

Reinforcement learning (RL) is another kind of ML. M. Ozturk et al. [17] balances power consumption and QoS in 5G UDHetNet. They dynamically switch small BSs based on the open call detail record (CDR) data set to reduce energy consumption while guaranteeing QoS. The results show that the performance of their algorithm is quite similar to the exhaustive search method. [8] proposes a branch of RL called Markov Decision Process (MDP) to solve energy-efficient BS control. Since in a real system, it is challenging to determine policy which is significant in RL, the authors design a Policy Rollout Algorithm to approximate the policy of MDP. Besides that, as the quantity of possible actions is quite large, it is impossible to calculate rewards of every action. To mitigate this drawback, they classify actions and rank them into different action spaces, actions inside the best action space and their rewards will be selected and calculated. This procedure is named Modification of Action Space. Results illustrate that energy can be saved by their algorithm. Different from the previous solution, in [18], Deep Neural Networks (DNN) is introduced to modify the classical Actor-Critic (AC) Reinforcement Learning (RL) framework because it cannot handle problems on a large scale. In their work, policy and interpreter, which are two parts of the agent, are approximated by DNN. In the training phase, they select a cutting-edge algorithm called the Deep Deterministic Policy Gradient (DDPG) training framework [19]. The result illustrates that the new reinforcement architecture can be executed efficiently even the quantity of small BSs is large. Yet, their algorithm is perfect, and they do not consider the spectrum leasing revenue of primary MNOs.

2.3 Spectrum leasing methods

L. Sboui et al. [20] provides a mathematical model of spectrum sharing and leasing between the primary network (PN) which is owned by the primary MNO, and secondary network (SN) which is owned by the secondary MNO. Furthermore, a simple algorithm is proposed to optimize energy consumption. Their mathematic model perfectly describes the spectrum leasing problem, but a more intelligent algorithm should be designed to boost energy saving. The same work is also finished in [21] by a different solution. A probability-based model of primary users (PU) and secondary users (SU) is established. A dynamic flow-adaptive leased spectrum strategy (DFSL) is designed to adjust the number of leased spectrums. The continuous-time Markov chain (CTMC) is also applied to analysis the model. Results demonstrate that the algorithm can enhance the service provided to SU. The same work is also researched by [22] and [23].

In [24], the paper provides a framework about dynamic spectrum access. They creatively applied a basic Upper-confidence Bound (UCB) method which is a branch of reinforcement learning to spectrum sensing and reusing. Because the environment of Cognitive Radio (CR) is complex, they simplified the environment into discrete time slots and proposed a Cognitive Agent (CA) which can interact with the environment optimally.

2.4 Contributions

The contribution of this project is fourfold:

1. An energy-saving and spectrum leasing combined mathematical model is proposed.

2. Epsilon greedy (E-greedy) method and upper confidence bounds (UCB) method of multi-armed bandit problem (MBP) is applied to maximize the revenue of primary MNOs and help secondary MNOs leasing spectrum.

3. Compare the performance of E-greedy and UCB to illustrate that UCB has the better performance whereas E-greedy cannot perform well.

4. A novel algorithm named clustering-based algorithm is proposed to cover the shortage of E-greedy and UCB.

3 Methodology

3.1 Reinforcement learning based solution

3.1.1 System Modelling of reinforcement learning

3.1.1.1 5G ultra-dense cellular network energy saving model

In 5G UDHetNet, five kinds of devices, including Macro Base-station (Macro BS), Micro Base-station (Micro BS), Pico Base-station (Pico BS), Femto Base-station (Femto BS), and remote radio head (RRH), can emit electromagnetic wave (EM wave) and utilize spectrum. Their relationship is shown in **Figure 1**. In the system, each kind of device will consume electric energy and emit EM waves and spectrum inside their own cell. Macro BS should take charge of the whole system and provide service for all users in Macro Cell, and other kinds of devices will offer services to a particular part of users.

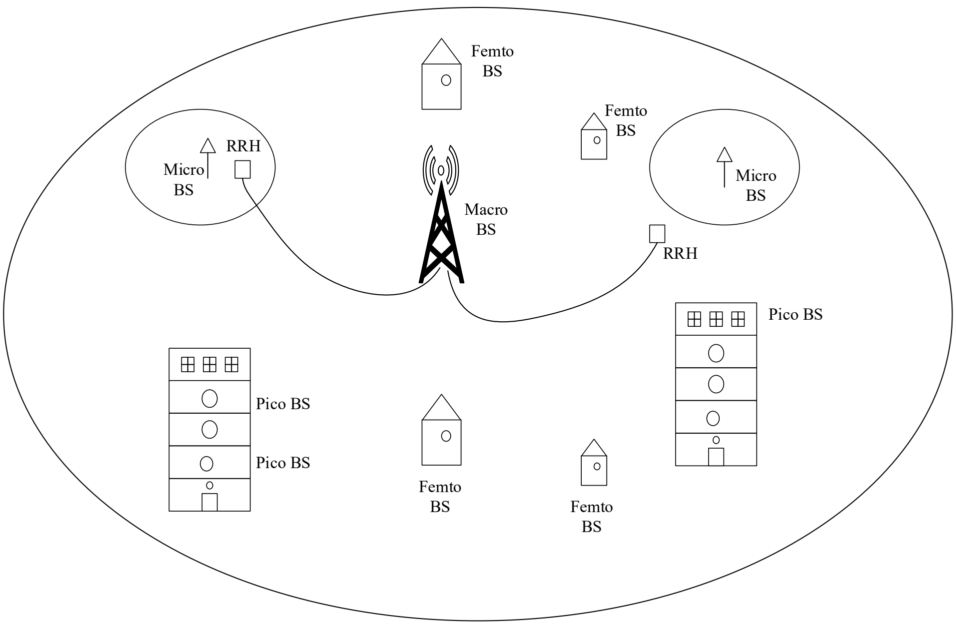

Figure 1: Diagram of a 5G UDHetNet cell, which contains five kinds of devices

The mathematical model and parameter matrix of UDN provided in [25] is suitable for system power calculation. The authors provided a linear approximated formula to calculate the power consumption of BSs. After modifying, the power of Macro BS is P_{Ma} which can be expressed as:

$$P_{Ma} = P_0^{Ma} + \Delta P^{Ma} \times \tau^{Ma} \times P_{Max}^{Ma} \quad watt \tag{1}$$

Where P_0^{Ma} is minimized power consumption of Macro BS, ΔP^{Ma} is the increasing slop, τ^{Ma} is the normalized traffic load and P_{Max}^{Ma} is the maximum power that a base station can output. These parameters are provided in the parameter matrix [25].

Similarly, power consumptions of other kinds of BSs can be expressed as:

$$P_i = \begin{cases} P_{0,i}^a + \Delta P_i^a \times \tau_i^a \times P_{Max,i}^a watt \ \ base\ station\ (a,i)\ is\ switched\ on \\ P_{sleep,i}^a\ watt \ \ base\ station\ (a,i)\ is\ switched\ off \end{cases} \quad (2)$$

Where a represents type of BSs, different types of BS, i is used to denote the index of BS and P_{sleep} represents the power consumption while the BS is switched off.

The overall system power consumption can be expressed as:

$$P_{total} = P_{Ma} + \sum_{i=1}^{end} P_i \text{ watt} \quad (3)$$

3.1.1.2 Spectrum Leasing model

Intuitively, the price of a unit of a commodity is determined by its cost and its demand. Furthermore, spectrum resources that belong to primary MNOs can be solen in terms of time slots. We can regard it as a spectrum source slot (SS). Thus, the overall revenue that a primary MNO can earn from spectrum leasing at a certain time can be formulated as:

$$R_t = \sum_{i=1}^{end} c_{t,i} \times \tau_{t,i} \times N_{t,i} \ \ dollors \quad (4)$$

Where t is the index of SS, c_i is the cost of the spectrum from BS i, τ_i is the normalized secondary MNOs' traffic demand at BS i. Finally, N is a constant that is used to modifies the price of SS to a reasonable level.

3.1.1.3 Problem formulation

The CDR data set from Milan city in [17] is applied as training and testing data set in this project. The set contains CDR data of 1 Macro BS and n small BSs during T time slots. Because all data is between 0 and 1, which can be treated as normalized traffic demand. Because the Macro BS cannot be switched off and every other small BS has 2 conditions (on/off), all combinations of BSs' states can be represented as A. **Table 1** illustrates the relationship between A and i, where i is the digits of A:

$$t \in \{0, 1, 2, \dots\dots, T\}$$
$$i \in \{1, 2, \dots\dots, n\} \quad (5)$$
$$A \in \{0, 1, 2, 3, 4, \dots\dots, 2^n - 1\}$$

A\i	1	2	…
0	0	0	0
1	1	0	0
2	0	1	0

Table 1: A-i mapping table. i is digits of A

Because those BSs who are switched on cannot off-load traffic and lease spectrum, a realistic spectrum leasing model can be modified from **(4)**:

$$R_t = \sum_{i=1}^{end} c_{t,i} \times \tau_{t,i} \times N_{t,i} \times \delta_{t,i} \quad dollors \quad (6)$$

Where $\delta_{t,i}$ is 0 when the BS i is "on" at t and 0 when the BS i is "off" at t.

The total revenue that a primary MNO can earn by selecting certain action **A** at certain time slot **t** can be represented as:

$$Rev_t(A) = \omega_1 \times R_{t,A} + \omega_2 \times P_{total,t,A} \text{ dollors} \quad (7)$$

Where $\boldsymbol{\omega_1}$ and $\boldsymbol{\omega_2}$ are weights of spectrum leasing revenue and energy-saving power respectively. $Rev_t(A)$ is the revenue of condition A at time slot t.

The target of this project is to find the condition (action) $A_{best,t}$ to guarantee $Rev_t(A)$ is maximized. This operation can be expressed as:

$$A_{best,t} = argmax_a\left(Rev_t(a)\right) \quad \forall a \in A \quad \tau_{t,i} \geq 1 - \tau_t^{Ma}$$

(8)

$1 - \tau_t^{Ma}$ is the traffic that a Macro BS can support at time slot t. $\tau_{t,i} \geq 1 - \tau_t^{Ma}$ guarantees that any small BS whose traffic demand is larger than the maximum bandwidth that a Macro BS can support must not be switched off. By this operation, the QoS of primary

MNO is kept. However, this target can be only achieved by the method of exhaustion (MoE) which is impossible in many practical situations. In reality, executing time is limited but MoE needs a large amount of time. Thus, MoE is not applicable and most works are done on artificial intelligence (AI) which is an alternative solution.

3.1.2 Reinforcement learning and multi-armed bandit problem

3.1.2.1 Reinforcement learning

Reinforcement learning (RL) is a part of artificial intelligence (AI) machine learning (ML) techniques. It is about mapping actions to rewards and select action so that an agent can maximize the reward [26]. The basic concepts of reinforcement learning are shown in **Figure 2**. An agent can interact with the environment by choosing actions. Once an action is selected, an interpreter can calculate reward and state according to the environment and the previous action. The reward and state will be sent back to the agent. During the next time slot, the agent can make a decision that may maximize the reward based on the previous rewards and states. After several loops, knowledges about the environment can be learnt by the agent by receiving rewards and states. The learning phase is called exploration and the phase where the actions are selected is exploitation. The main challenge of reinforcement learning is balancing exploration and exploitation. A great algorithm should on the one hand guarantee reward while, on the other hand maintains exploring new opportunities. The Multi-armed bandit problem (MBP) is a classical scenario of reinforcement learning. Epsilon greedy and upper confidence bound are two common methods to solve MBP.

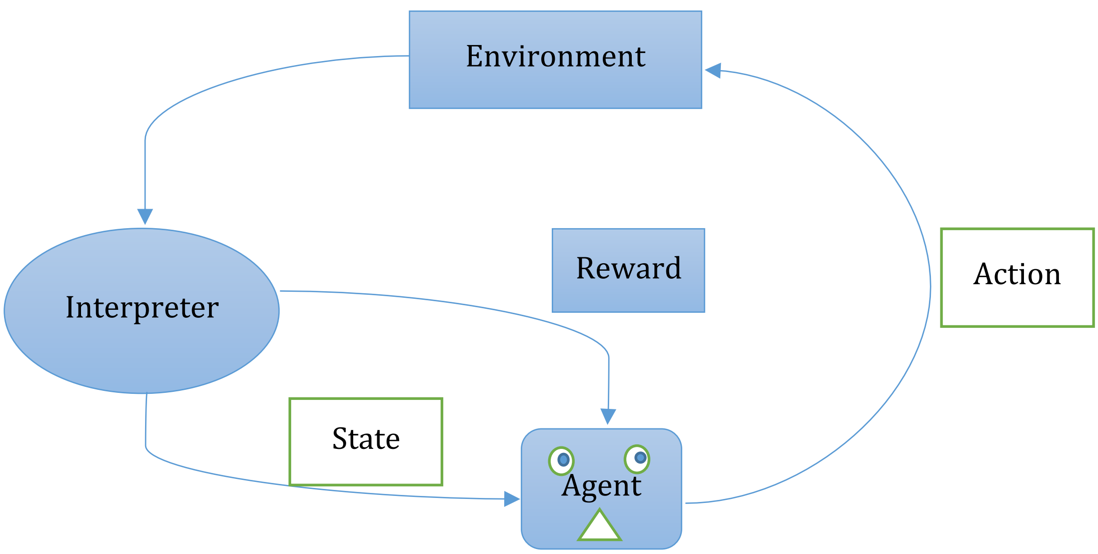

Figure 2: Diagram of reinforcement learning

3.1.2.2 Epsilon greedy method of energy saving and spectrum leasing

Epsilon greedy (E-greedy) method is the most common and simplest method of MBP [26]. A parameter called epsilon is used to determine the probability of exploring an imperfect action. The precise procedure is shown in **Algorithm 1** where t is the number of time slots, Q is an array whose length is the same as the total quantity of actions that is valid to be chosen, R is reward table whose length is equal to Q, A is an action that that chosen at a certain time slot, epsilon is a parameter belongs to [0, 1], Random () is a function that can generate random number between its input, N is an array that records how much times an action has been chosen. Q-table of each action can be calculated by **Equation 1** which is called an action-value estimate. The general form of **Equation 1** is given by [27] where $[Target - OldEstimation]$ is called error:

$$NewEstimation \leftarrow Oldestimation + StepSize\,[Target - OldEstimation]$$

The error will adjust the previous estimation to a more credible value once an action is executed.

Obviously, if t is large enough, **Algorithm 1** will finally explore all actions and make a good decision based on Q table. Lower epsilon will eventually cause better performance because the algorithm has a higher probability of choosing the best action. On the contrary, a higher epsilon value can boost exploration and shorten convergency time due to the fact that the algorithm prefers to choose unknown actions. This preference can help the agent understand the whole environment quickly.

```
Algorithm 1: Typical E-greedy method
Q = 0
R = 0
A = 0
N = 0
For t from 1 to final:
    If Random (0, 1) > Epsilon:
        If t == 1:
            A = Random (1, 2^n − 1)
            N[A] = N[A] +1
                Calculate Rev[A]
        Else:
            A = argmax_a (Q)
            N[A] = N[A] + 1
            Calculate Rev[A]
        End if
    Else:
        A = Random (1, Total number of actions)
        N[A] = N[A] +1
        Calculate Rev[A]
    End if
    Q[A] = Q[A] + (Rev[A] - Q[A]) / N[A]                Equation 1
    Output: A
End for
```

Algorithm 2: Procedure of Typical UCB method

lr = learning rate
Q = 0
A =0
N = 0
Rev = 0
n = total quantity of small BSs
T = total quantity of time slots

For t in range $(1,\ \ 2^n - 1)$
    A = t
    $\mathrm{Q[A]} = \mathrm{Q[A]} + lr \times (Rev_t(A)\ -\ \mathrm{Q[A]})$
End for
For t in range ($2^n - 1$, T)
    $A\ =\ argmax_A(Q\,(A)\ +\ c\sqrt{\frac{lnt}{N\,(A)}})$
    $\mathrm{Q[A]} = \mathrm{Q[A]} + lr \times (Rev\,(A)\ -\ \mathrm{Q[A]})$
    Output: A
End for

3.1.2.3 Upper-confidence Bound (UCB) Method of energy saving and spectrum leasing

UCB is another method to solve MBP. Different from the E-greedy method, the UCB method does not adopt any random number. It only depends on the training data. **(9)** illustrates how an action is selected:

$$A_t\ =\ argmax_A(Q_t(A)\ +\ c\sqrt{\frac{lnt}{N_t(A)}}) \tag{9}$$

t is the index of current time slot, $N\,(A)$ is the times that action A has been selected, c is a parameter whose value is larger than 0. If N(A) is low, and t is large. In this circumstance, action A is probably being selected. (9) guarantees that the agent will explore enough actions before the exploitation phase. A simplified process of UCB is shown in **Algorithm 2**.

In **Algorithm 2**, learning rate is a parameter that is set by engineers. This parameter defines the weight of the latest error in Q[A]. Higher learning rate will update Q faster. Rev[A] which can be calculated by **(7)** is the revenue of action A.

3.1.2.4 Modified MRB algorithm:

```
Algorithm 3: Modified UCB method
lr = learning rate
PP = punishment parameter
TH = threshold
Q = 0
A = 0
N = 0
TD = training duration
Q = 0
A = 0
N = 0

#Training:
        For t in range (1, TD)
                For all actions A
                        If Rev_t(A) ≤ TH
                                Q[A] = PP × Q[A]
                        Else
                                Q[A] = Q[A] + lr × (Rev_t(A) − Q[A])
                        End if
                End for
        End for
#Testing:
        For t in range (TD, T)
        A_t = argmax_A(Q_t(A) + c sqrt(lnt / N_t(A)))
                Q[A] = Q[A] + lr × (Rev_t(A) − Q[A])
        Output: A
        End for
```

Unfortunately, **Algorithm 1** and **2** are only available when the quantity of t is much larger than the quantity of A. In many practical conditions, exploring all actions one by one needs a large amount of time. Thus, a modified algorithm should be designed. To help UCB and E-greedy converge faster, the typical E-greedy or UCB algorithms should be divided into two parts: training and testing. In the training phase, the algorithm will calculate the rewards of all actions to generate a Q table. This operation can rapidly converge UCB and E-greedy.

The modified edition of UCB is illustrated in **Algorithm 3**. Two parameters called punishment parameter (PP) whose value is between 0 to 1, and threshold (TH) which is used to ensure QoS are proposed. During the training phase, rewards and Q values of all actions should be calculated and recorded. Because in small BSs dynamic control, the priority is QoS, any action that hazards QoS should be punished. For this reason, any action whose reward is less than TH should be multiplied with PP whose value is between 0 to 1. Lower PP value will cause more serious punishment. Data between $t \in [1, TD]$ is chosen as the training data set and $t \in [TD + 1, 1008]$ is the testing data set.

3.2 Clustering-based method of energy saving and spectrum leasing

3.2.1 Clustering-based method of 5G system

Clustering is used to label data [28] or classifies objects. Clustering contains many good performed and complex methods including k-means, Agglomerative, etc. [29]. However, because the research of applying clustering into switching off BSs is in its infant, a pretty simple clustering algorithm is designed in this project.

The basic concept of the clustering-based method is quite simple. In a 5G UDN system, those small BSs who obtain higher traffic demand must not be switched off. For this reason, finding a set of small BSs whose traffic demand is low and idling them is a potential solution.

Figure 3 demonstrates what is clusters of small BSs in a 5G system.

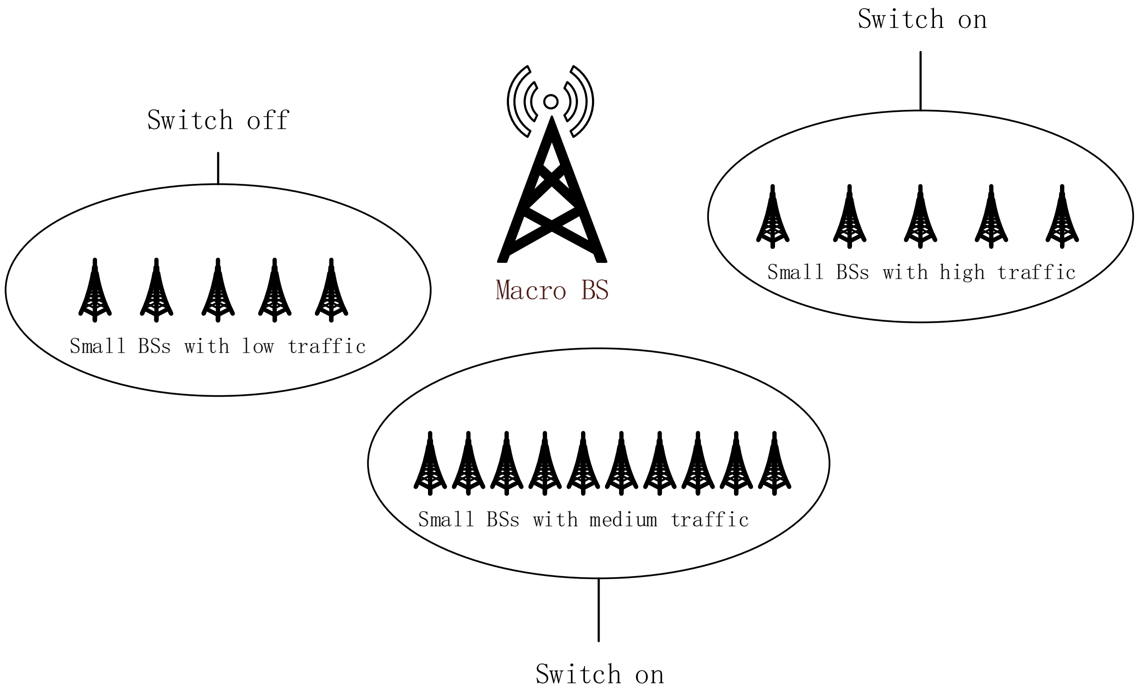

Figure 3: Classifying small BSs into clusters.

3.2.2 Procedure of Clustering-based method

In **Algorithm 4,** firstly, those small BSs whose traffic demand is greater than the remaining traffic that a Macro BS can support should not be idled because once these small BSs are switched off, QoS cannot be guaranteed. In the next step, the remaining small BSs will be ranked according to their traffic demand. The idea is to classify the ranked BSs into several clusters so that the cluster with the smallest traffic load can be switched off. This process can be simplified as **Equation 2**. The target of **Equation 2** is finding the largest small BS set that can occupy the highest traffic of Macro BS. For example, the remaining traffic of Macro BS in **Table 2** is 0.8. The sum of the previous three small BSs is 0.7, which is slightly smaller than Macro BS. However, if the fourth small BS is added, the sum of traffic demand becomes 1.1, which is larger than 0.8. Thus, only small BSs 1-3 can be switched off.

The clustering method can inherently guarantee QoS because it is based on traffic load. Its executing duration is significantly lower than MBP, especially when the quantity of small BSs in a system is high. However, power-saving and spectrum leasing are by-products and are not the central concerned part of this algorithm.

Switch off

Traffic SBS / Remain Traffic MBS	1	2	3	4	5
0.8	0.15	0.2	0.35	0.4	0.7

0.7

Table 2: Example of procedure of Clustering method

```
Algorithm 4: Cluster-based switching off method

TMa[t] = traffic of Macro BS by CDR at t
TSm[i, t] = traffic of small BS i by CDR at t
Traffic_Sum = 0

For t in range (1, final)
        If 1 – TSm[i, t] > TMa[t]
                Delete small BS i from TSm
        Else
                Rank small BS i based on TSm[i, t]
        End if

For i in range (1, final)
                Traffic_Sum = Traffic_Sum + TSm[i, t]             Equation 2
                Record i as array Switch_off_set
                If Traffic_Sum ≥ TMa[t]
                        Remove i from Switch_off_set
                        Break
                End if
        End for

Switch off the Switch_off_set
End for
```

4 Results

4.1 Results of MBP

The performance of an algorithm is described by the ratio of selected action's reward and the best action's reward. Mathematically, it can be written as:

$$Perf[t] = \frac{reward\ of\ the\ selected\ action\ at\ t}{reward\ of\ the\ best\ cation\ at\ t} \quad \forall\ t \in \{1, 2, 3, \dots \dots, T\} \tag{10}$$

The performance of algorithms is evaluated by the ratio between the reward of the selected action and the reward of the best action. If Perf equals to 1, that means the best action is selected. If Perf equals 0, no reward can be earned from this action. Perf is blank when an action cannot guarantee QoS.

The CDR data in [12] is applied as training and testing data set of MBP. This data set contains 12 small BSs and a Macro BS. It records CDR data of Milab city every 10 minutes for a week. Epsilon of E-greedy method is 0.5. Data from $t \in [1, 100]$ is used to train MBP. The remaining data is used to test MBP.

4.1.1 Effect of training:

The result provided by [24] cannot converge within a small-time duration. Similarly, without the training phase, the UCB method cannot converge within 1008 time slots. **Figure 4** and **Figure 5** demonstrate the comparison of UCB with and without the training. It is clear that the performance of UCB without training is quite bad because cannot find actions that can guarantee QoS, whereas UCB with training performs well. Especially, **Figure 4** illustrates that UCB without training cannot choose any action that does not deter QoS between $t \in [1, 850]$. The reason of this phenomenon is that UCB needs a long time to explore the environment. Oppositely, the majority of selected actions in **Figure 5** can not only guarantee QoS but also maximize revenue.

4.1.2 Comparison between E-greedy with training and UCB with training:

From **Figure 5** and **Figure 6**, it is evident that the performance of UCB is much better than E-greedy. The reason for this phenomenon is that the E-greedy method is too simple to predict the 5G system. Because more than 87% of actions cannot guarantee QoS, it will probably reach a low performance if choosing actions randomly.

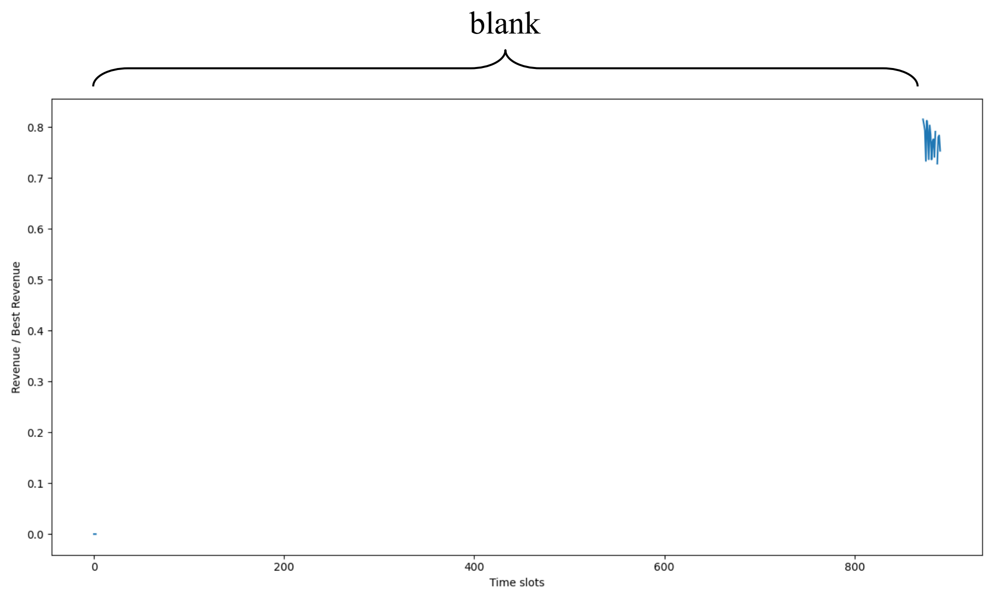

Figure 4: Performance of UCB without training. Perf value before time slot = 850 is blank, which means actions selected in these time slots cannot guarantee QoS.

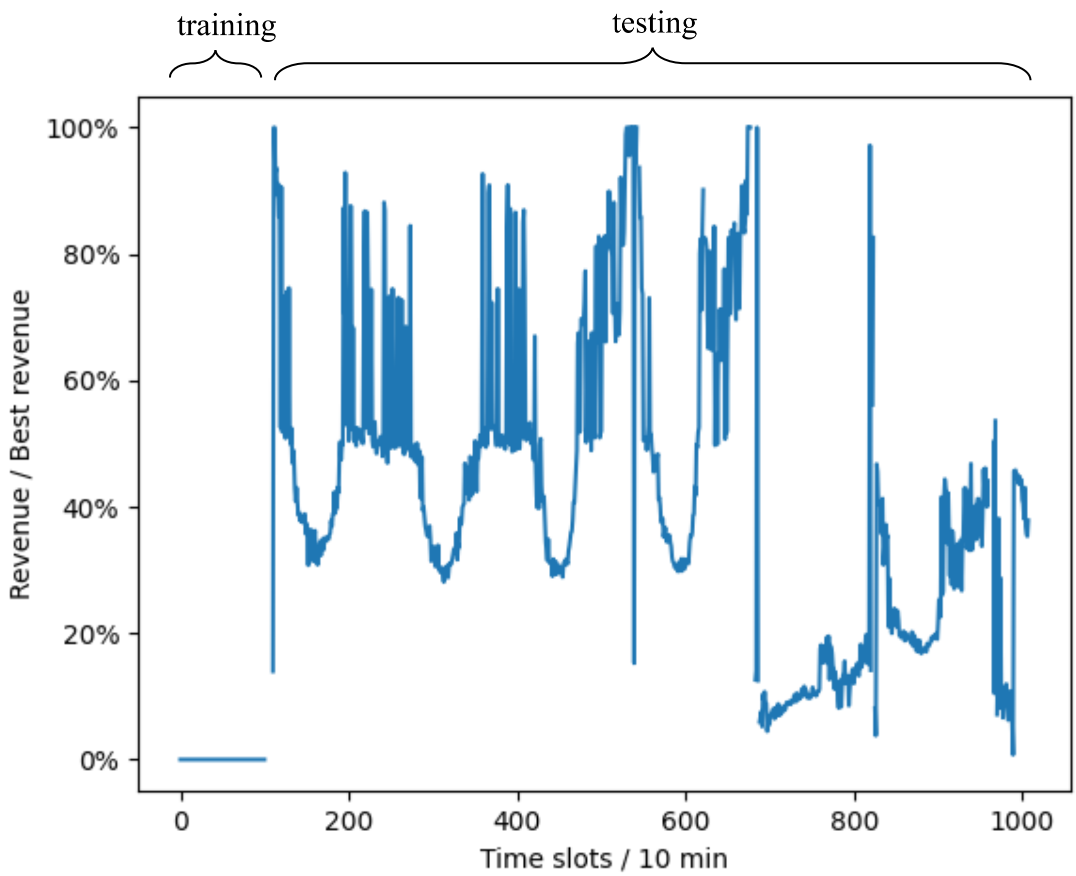

Figure 5: Performance of UCB with training. The majority of time slots have value. The average value of Perf is 45%, which is a good performance.

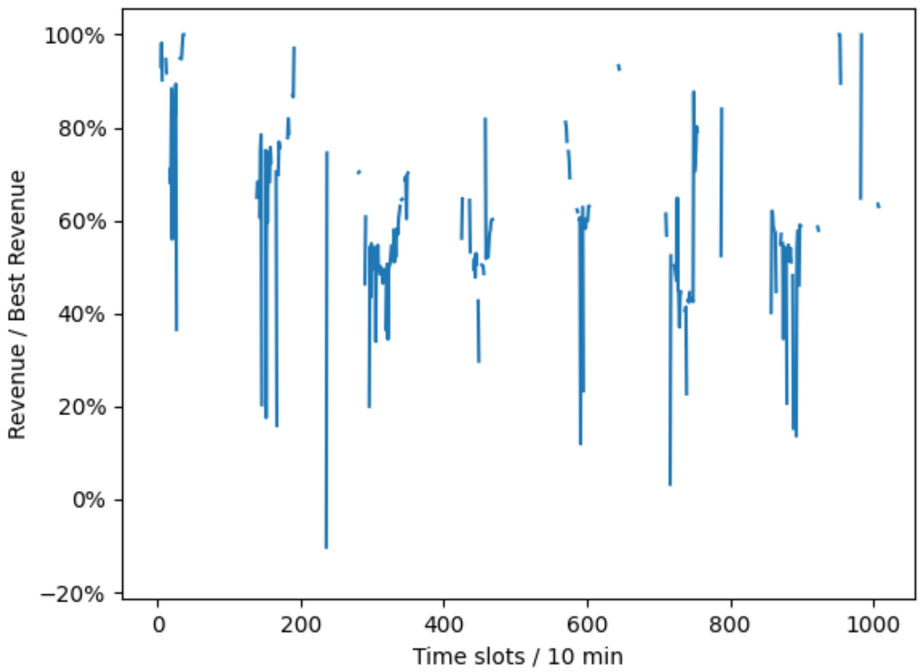

Figure 6: Performance of E-greedy method with training. Some time slots have value. However, E-greedy method is too simple to get a good performance.

4.1.3 Limitations of MBP

Although MBP can maximize spectrum leasing and energy saving revenue in a good manner, it has a significant drawback. For a system which contains 101 BSs, it is impossible to train MBP algorithm and update Q-table. Compared with clustering method, MBP needs more calculation.

4.2 Results of Clustering-based method:

Similar to the CDR data set that has been applied in MBP, another CDR data set that contains 1 Macro BS and 100 small BSs is used in this simulation. By applying **(3)** and **(4)**, the result is generated.

4.2.1 Energy saving and spectrum leasing revenue

The performance of the clustering-based switching method can be illustrated in **Figure 7, 8, 9**. As shown in **Figure 7**, 0.46% of total energy is saved by this algorithm averagely, which is relatively low. The reason for this phenomenon is that there is not too much remaining traffic of Macro BS. As a result, only a few small BSs can be switched off. In **Figure 8**, the revenue of spectrum leasing is 28 dollars in average.

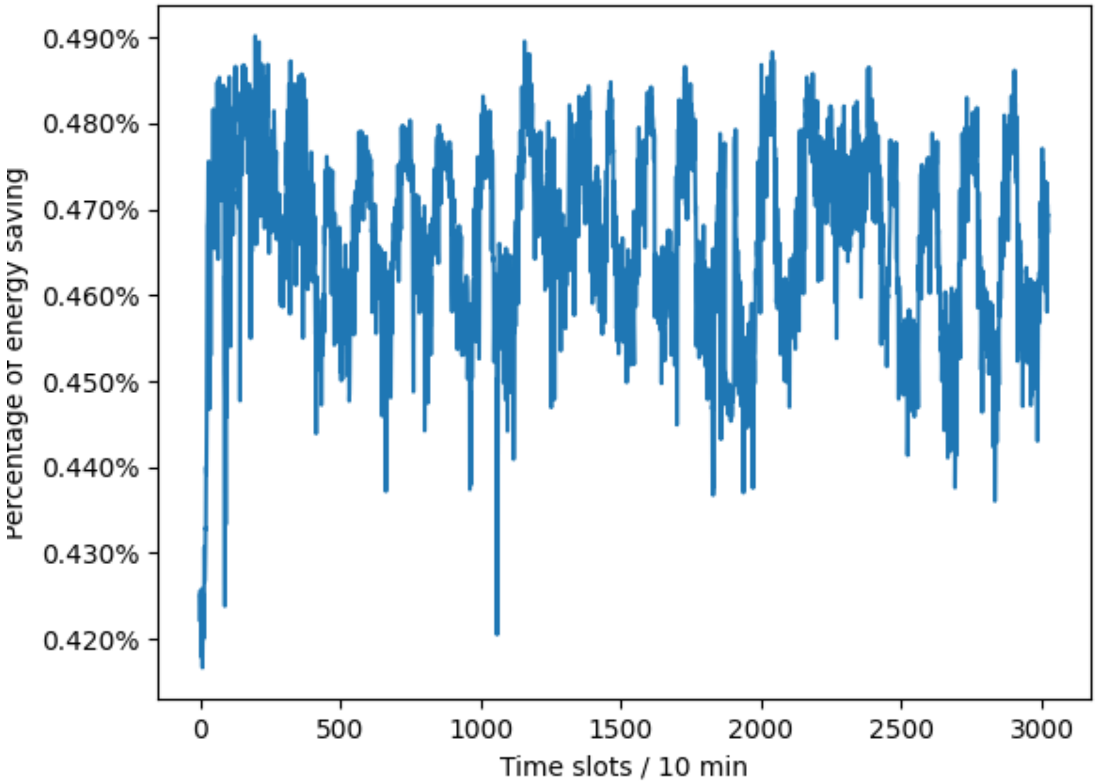

Figure 7: Percentage of energy saved by clustering method

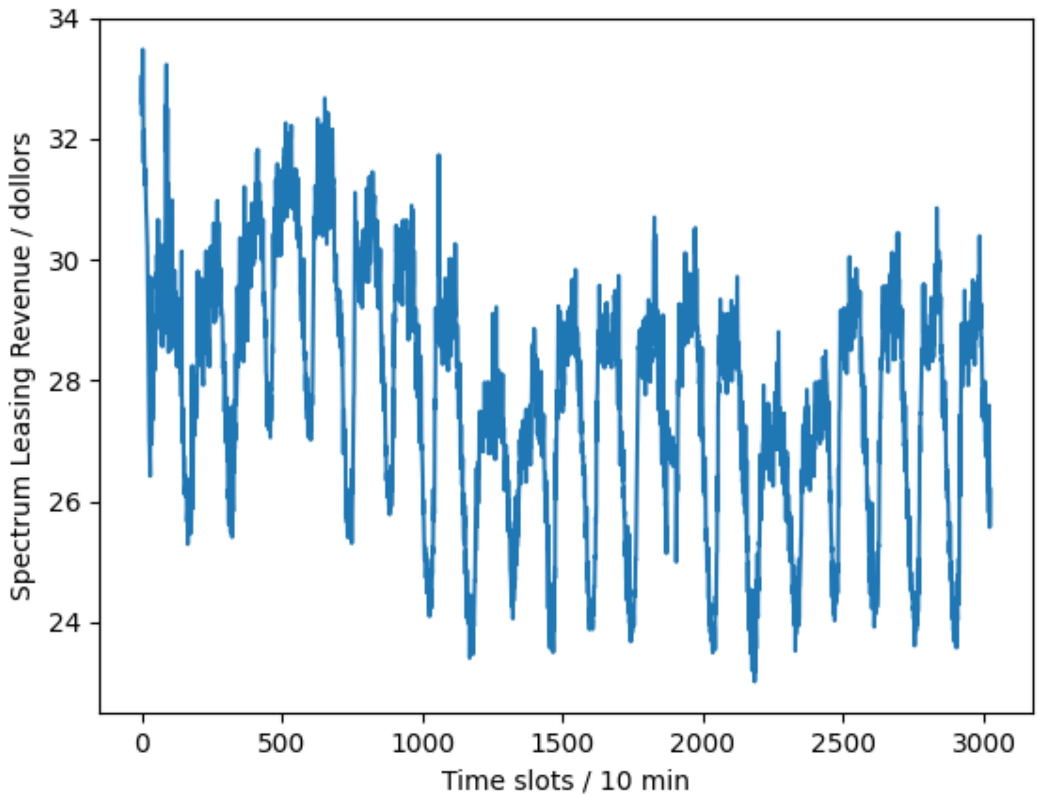

Figure 8: Spectrum leasing revenue of clustering method

4.2.2 Overall performance of Clustering-based method

In **Figure 9**, the overall revenue of energy-saving and spectrum leasing is shown. According to **(7)**, ω_1 is 0.2 and ω_2 is 0.5. Unfortunately, because it is impossible to train E-greedy, UCB, or calculating the highest revenue by this data set, we cannot compare the results of clustering-based method with MBP or the best solution.

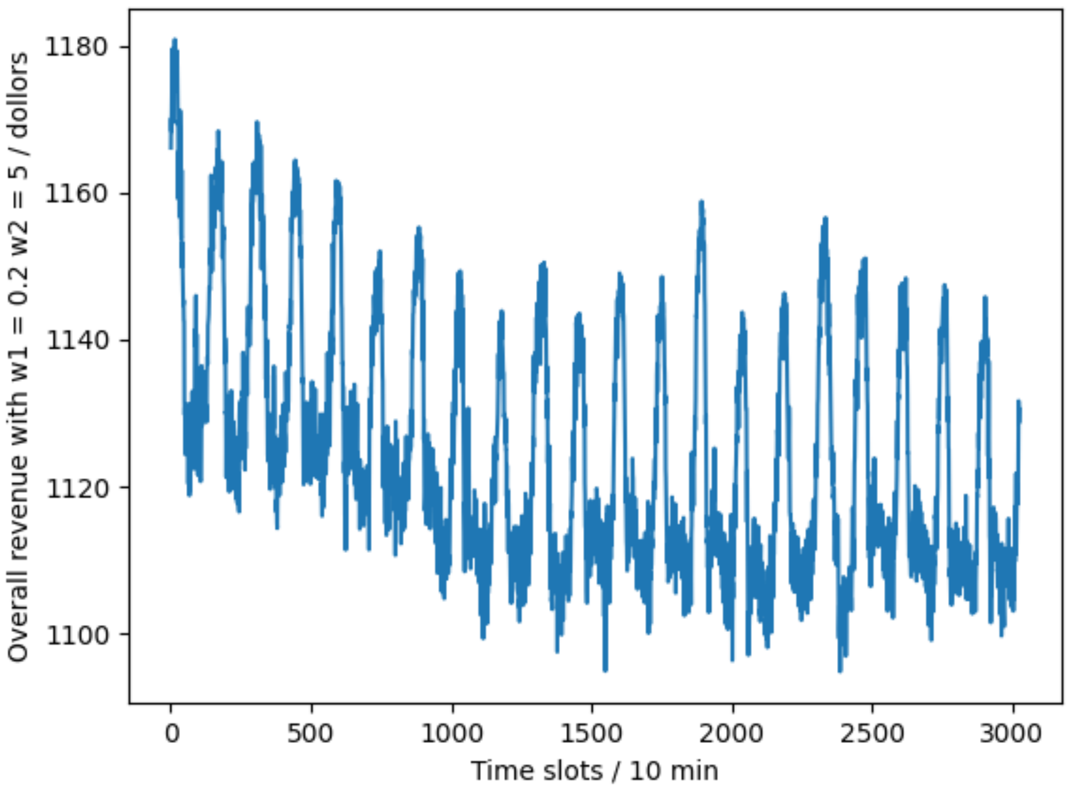

Figure 9: Overall revenue of clustering.

4.2.3 Limitations of clustering-based method

Clustering can boost executing and save time. It can be applied on a system contains a large number of BS. However, the clustering method only concentrates on the traffic demand of BSs. Energy saving and spectrum leasing revenue is not the priority. Thus, it is hard for a clustering method to find a good switching on/off pattern.

5 Conclusion:

MBP in RL can solve spectrum leasing and energy saving problem perfectly. A CDR data set of 5G system with 12 small BSs and a Macro BS is used to test MBP algorithm. Compared with the highest revenue, one branch of MBP called UCB can reach around 50% of the best performance and reduce calculating time duration dramatically. However, without training phase, UCB or E-greedy cannot converge quickly.

MBP cannot solve a system contains 100 small BSs. Thus, the clustering-based method is proposed to solve this problem. Results show that although the performance of clustering method is not as good as MBP, this algorithm can be applied in a system containing 101 BSs and saving energy, earning spectrum leasing revenue for primary MNOs.

6 References

[1] J. G. Andrews et al., "What Will 5G Be?," in IEEE Journal on Selected Areas in Communications, vol. 32, no. 6, pp. 1065-1082, June 2014, doi: 10.1109/JSAC.2014.2328098.

[2] X. Ge, S. Tu, G. Mao, C. Wang and T. Han, "5G Ultra-Dense Cellular Networks," in IEEE Wireless Communications, vol. 23, no. 1, pp. 72-79, February 2016, doi: 10.1109/MWC.2016.7422408.

[3] F. Elsherif, E. K. P. Chong and J. -H. Kim, "Energy-Efficient Base Station Control Framework for 5G Cellular Networks Based on Markov Decision Process," in *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 9267-9279, Sept. 2019, doi: 10.1109/TVT.2019.2931304.

[4] P. Gandotra, R. K. Jha, and S. Jain, "Green communication in next generation cellular networks: A survey," IEEE Access, vol. 5, pp. 11 727– 11 758, 2017.

[5] K. Chang, K. Chu, H. Wang, Y. Lin and J. Pan, "Energy Saving Technology of 5G Base Station Based on Internet of Things Collaborative Control," in IEEE Access, vol. 8, pp. 32935-32946, 2020, doi: 10.1109/ACCESS.2020.2973648.

[6] K. N. R. S. V. Prasad, E. Hossain and V. K. Bhargava, "Energy Efficiency in Massive MIMO-Based 5G Networks: Opportunities and Challenges," in IEEE Wireless Communications, vol. 24, no. 3, pp. 86-94, June 2017, doi: 10.1109/MWC.2016.1500374WC.

[7] I, CL., Han, S. & Bian, S. Energy-efficient 5G for a greener future. Nat Electron 3, 182–184 (2020). https://doi.org/10.1038/s41928-020-0404-1

[8] M. Feng, S. Mao and T. Jiang, "Base Station ON-OFF Switching in 5G Wireless Networks: Approaches and Challenges," in IEEE Wireless Communications, vol. 24, no. 4, pp. 46-54, Aug. 2017, doi: 10.1109/MWC.2017.1600353.

[9] F. Richter, A. J. Fehske and G. P. Fettweis, "Energy Efficiency Aspects of Base Station Deployment Strategies for Cellular Networks," *2009 IEEE 70th Vehicular Technology Conference Fall*, Anchorage, AK, USA, 2009, pp. 1-5, doi: 10.1109/VETECF.2009.5379031.

[10] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, ''Toward dynamic energyefficient operation of cellular network infrastructure,'' IEEE Commun. Mag., vol. 49, no. 6, pp. 56–61, Jun. 2011.

[11] A. Mohamed, O. Onireti, M. A. Imran, A. Imran and R. Tafazolli, "Control-Data Separation Architecture for Cellular Radio Access Networks: A Survey and Outlook," in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 446-465, Firstquarter 2016, doi: 10.1109/COMST.2015.2451514.

[12] K. A. Yau, J. Qadir, C. Wu, M. A. Imran, and M. H. Ling, “Cognition-inspired 5g cellular networks: A review and the road ahead,” IEEE Access, vol. 6, pp. 35 072–35 090, 2018

[13] F. Hussain, S. A. Hassan, R. Hussain and E. Hossain, "Machine Learning for Resource Management in Cellular and IoT Networks: Potentials, Current Solutions, and Open Challenges," in *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1251-1275, Secondquarter 2020, doi: 10.1109/COMST.2020.2964534.

[14] Alsafasfeh, Q.; Saraereh, O.A.; Ali, A.; Al-Tarawneh, L.; Khan, I.; Silva, A. Efficient Power Control Framework for Small-Cell Heterogeneous Networks. Sensors 2020, 20, 1467. https://doi.org/10.3390/s20051467

[15] B. Shen, Z. Lei, X. Huang and Q. Chen, "An Interference Contribution Rate Based Small Cells On/Off Switching Algorithm for 5G Dense Heterogeneous Networks," in *IEEE Access*, vol. 6, pp. 29757-29769, 2018, doi: 10.1109/ACCESS.2018.2841044

[16] Y. Li, H. Zhang, J. Wang, B. Cao, Q. Liu and M. Daneshmand, "Energy-Efficient Deployment and Adaptive Sleeping in Heterogeneous Cellular Networks," in *IEEE Access*, vol. 7, pp. 35838-35850, 2019, doi: 10.1109/ACCESS.2019.2892226.

[17] M. Ozturk, A. I. Abubakar, J. P. B. Nadas, R. N. B. Rais, S. Hussain and M. A. Imran, "Energy Optimization in Ultra-Dense Radio Access Networks via Traffic-Aware Cell Switching," in *IEEE Transactions on Green Communications and Networking*, doi: 10.1109/TGCN.2021.3056235.

[18] J. Ye and Y. J. A. Zhang, “Drag: Deep reinforcement learning based base station activation in heterogeneous networks,” IEEE Transactions on Mobile Computing, vol. 19, no. 9, pp. 2076–2087, 2020.

[19] T. P. Lillicrap, J. J. Hunt, A. Pritzel, et al., “Continuous control with deep reinforcement learning,” in International Conference on Learning Representations, San Juan, Puerto Rico, May 2016.

[20] L. Sboui, H. Ghazzai, Z. Rezki and M. Alouini, "Green collaboration in cognitive radio cellular networks with roaming and spectrum trading," *2015 IEEE 26th Annual*

International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong, China, 2015, pp. 1420-1425, doi: 10.1109/PIMRC.2015.7343521.

[21] Xiao X, Zeng F, Hu Z, Jiao L. Dynamic Flow-Adaptive Spectrum Leasing with Channel Aggregation in Cognitive Radio Networks. Sensors (Basel). 2020 Jul 7;20(13):3800. doi: 10.3390/s20133800. PMID: 32645964; PMCID: PMC7374385.

[22] Fan, R., Chen, W., Jiang, H., An, J., Yang, K., & Xing, C. (2018). Dynamic spectrum leasing with two sellers. *IEEE Transactions on Vehicular Technology*, *67*(6), 4852-4866.

[23] Handouf, S., Sabir, E., & Sadik, M. (2018, October). A Leasing-based Spectrum Sharing Framework for Cognitive Radio Networks. In *2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM)* (pp. 1-5). IEEE.

[24] W. Jouini, D. Ernst, C. Moy and J. Palicot, "Upper Confidence Bound Based Decision Making Strategies and Dynamic Spectrum Access," *2010 IEEE International Conference on Communications*, Cape Town, South Africa, 2010, pp. 1-5, doi: 10.1109/ICC.2010.5502014.

[25] G. Auer *et al*., "How much energy is needed to run a wireless network?," in *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40-49, October 2011, doi: 10.1109/MWC.2011.6056691.

[26] Vermorel J., Mohri M. (2005) Multi-armed Bandit Algorithms and Empirical Evaluation. In: Gama J., Camacho R., Brazdil P.B., Jorge A.M., Torgo L. (eds) Machine Learning: ECML 2005. ECML 2005. Lecture Notes in Computer Science, vol 3720. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11564096_42

[27] Richard S. Sutton, and Andrew G. Barto, "Reinforcement Learning: An Introduction second edition", The MIT Press, Cambridge, Massachusetts.

[28] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1), e0210236.

[29] Rai, P., & Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, *7*(12), 1-5.