文章编号:2095-6134(2022)03-403-07

# 无线密集网络中的低损耗多臂老虎机算法*

赵耀[1,2,3],罗喜良[1†]

(1 上海科技大学信息科学与技术学院, 上海 201210; 2 中国科学院上海微系统与信息技术研究所, 上海 200050;
3 中国科学院大学, 北京 100049)

(2020 年 1 月 14 日收稿; 2020 年 4 月 20 日收修改稿)

**Zhao Y, Luo X L. A low cost multi-armed bandit algorithm for dense wireless network[J]. Journal of University of Chinese Academy of Sciences,2022,39(3):403-409. DOI:10.7523/j.ucas.2020.0011.**

**摘 要** 近年来人们对移动无线服务的需求与日俱增,为应对这一挑战,超密集无线网络被认为是下一代无线通信网络的基础设施架构和重要组成部分,基站的密集布置可以减少每个小区的服务用户数量,从而可为网络用户提供高速且低延迟的无线服务。但同时带来的不可避免的问题是用户在选择接入时会触发频繁的网络切换以确保可以接入到服务最佳的网络。用户接入问题往往被建模成在线学习模型。本文旨在寻找一个高效的在线用户接入方案以应对频繁网络切换造成的额外性能损失。通过对多臂老虎机模型的分析,提出基于操作杆淘汰机制的改进算法,并通过严格理论分析及数值仿真实验两个角度论证该算法的有效性。

**关键词** 在线学习;用户接入;密集网络;多臂老虎机

**中图分类号**:TN913 **文献标志码**:A **DOI**:10.7523/j.ucas.2020.0011

# A low cost multi-armed bandit algorithm for dense wireless network

ZHAO Yao[1,2,3], LUO Xiliang[1]

(1 *School of Information Science and Technology*, *ShanghaiTech University*, *Shanghai* 201210, *China*;
2 *Shanghai Institute of Microsystem and Information Technology*, *Chinese Academy of Sciences*, *Shanghai* 200050, *China*;
3 *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

**Abstract** In recent years, people's demand for mobile wireless services has been increasing. In order to meet this challenge, ultra-dense wireless networks are considered to be the infrastructure and important components of the next-generation wireless communication network. Massive deployment of small base stations can reduce the number of network users in each cell, which can in turn provide the users with high-speed and low-latency wireless service. However, the inevitable problem brought with it at the same time is that users will cause frequent network handover when choosing access to ensure that they can access the network with the best service provider. User association problem is often modeled as the online learning model. This paper aims to find an efficient online user association scheme to deal with the additional network performance loss caused by frequent handover. Based on the analysis of the multi-armed bandit (MAB) model, this paper proposes an improved algorithm based on the arm elimination strategy, and demonstrates the

* 国家自然科学基金(61971286)资助

† 通信作者,E-mail: luoxl@shanghaitech.edu.cn

effectiveness of the algorithm through rigorous theoretical analysis and numerical simulation experiments.

**Keywords** online learning; user association; dense network; multi-armed bandit

异构网络的快速发展在各类无线网络应用中发挥中极其重要的作用,例如,物联网、车联网、智慧工厂等。而为了应对正在快速增长的无线数据量的需求,无线网络密集化,即密集组网被认为是下一代无线网络发展的必然趋势之一[1]。各种类型小型基站,例如,微基站、皮基站、飞基站的密集布置对于扩大小区的覆盖范围以及增强用户基站之间链接的通信质量大有裨益,而相比较于只有宏基站部署的解决方案,各种小型基站的密集部署也有利于不同区域的频谱复用。

但是无线网络的密集化往往会导致严重的网络频繁切换的问题,有时也被称之为网络的乒乓效应,其会严重影响到网络性能,这也是目前密集无线网络面临的主要挑战之一。

为解决这一问题,近年来已有的很多研究工作把目光集中在自组织网络(self-organizing network, SON)中,并且这也是最早在 3GPP Rel-8 标准中提出的解决方案。

Lee 和 Cho[2]提出一个用户集体网络切换的解决方案,该方案通过优化用户接入初期阶段的延迟来降低网络中断的概率。Fischionei 等[3]提出混合型的系统模型来优化用户接入,并且该项工作同时考虑了用户的移动速度以及用户的距离造成的影响。文献[4]提出一个马尔科夫分析模型,利用特定的场景参数,例如路径损耗参数、用户移动速度,以及小区负载等等,最终得到依场景变化的最优网络切换解决方案。Ye 等[5]提出一个分布式负载感知的用户接入算法,该算法复杂度低且考虑了负载均衡以及用户接入过程中涉及的频谱划分等因素。文献[6]从降低能量消耗的角度出发,优化在用户接入过程中的频谱分配及功率分配。孟庆民等[7]提出一种基于马尔科夫链预测的切换方案,在触发网络切换时会预测下一个基站的相关状态信息。

但是以上的这些工作其实都很少考虑用户频繁切换网络的这个问题。在相关的参考文献当中,想尝试解决该问题更多地还是依赖于机器学习的方法。例如文献[8]借助于深度强化学习训练得到一个最优的系统控制器,该控制器可以用来降低网络切换的次数同时也保障了网络的吞吐量性能。目前为止和本文工作最相关的是文献[9],在该项工作当中,作者首次论述建立无线网络切换过程和多臂老虎机模型之间的等价性,除此以外该文献也提出一种改进的置信区间上界(upper-confidence bound, UCB)策略以较低复杂度来解决用户接入频繁切换的问题。

## 1 系统模型

### 1.1 网络模型

考虑一个区域范围内的无线密集网络,如图 1 所示,该区域内部署了前文所述的各类型基站:宏基站、微基站、皮基站、飞基站。这些基站的集合表示为

$$B = \{1,2,\cdots,N\}, \tag{1}$$

其中 $N$ 为基站总数目。用户的集合表示为

$$U = \{1,2,\cdots,K\}, \tag{2}$$

所有这些基站都可以为其覆盖范围内的用户提供无线服务。

为描述简洁,本文假设时间尺度划分为均等长度的时隙。在传统的无线网络系统中,用户接入与切换往往是由配备在基站端的控制器来决定的。但是很多研究已经表明在下一代无线网络系统中,由于网络的高度异构性,网络用户主导的接入方案正成为主流的趋势[1,10],这也是本文中所考虑的用户接入机制模型。

对于某一个特定的网络用户,记该用户在时隙 $t$ 所选择建立链接的基站为 $a_t \in B$。那么在总时间长度 $T$ 范围内的用户接入模式可以表示为

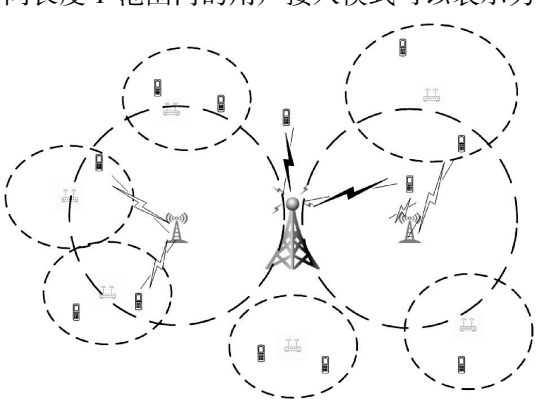

**图 1 无线密集网络示意图**

Fig. 1 Wireless dense network

$$\boldsymbol{\Theta} = (a_1, a_2, \cdots, a_T). \tag{3}$$

本文并不考虑其他复杂的接入模式,例如,3GPP Rel-12 中的双连接,即一个用户可以同时接入 2 个不同的基站接受服务。

用户在时隙 $t$ 结束后会立即收到奖励 $u(t, a_t)$。在本文中,我们旨在建立适用于用户接入模型的通用框架和解决思路,故在此并不指定这里奖励的具体指标。在实际的系统当中,这里的奖励通常刻画的是某项网络性能指标的优化,例如,最大化数据吞吐量、最小化网络时延,或者是最小化用户能量消耗等。

假设奖励由一个随机变量产生,出于理论分析的方便,这里将奖励归一化至 0~1。每一个基站都有各自的随机变量,其数学期望表达为

$$u_i = E[u(t,i)], i = 1, \cdots, N, \tag{4}$$

为方便表达同时也不失一般性,令这里的 $N$ 个数学期望是依升序排列

$$u_1 \leqslant u_2 \leqslant \cdots \leqslant u_N, \tag{5}$$

同时记它们之间的均值差距为

$$\Delta_i := u_N - u_i, i = 1, \cdots, N-1, \tag{6}$$

从时隙 1 开始计时至时隙 $T$ 为止,该用户所接受到的总奖励可以表示为

$$A(\boldsymbol{\Theta}) = \sum_{t=1}^{T} u(t, a_t), \tag{7}$$

用户的目标是在缺乏关于可向其提供无线服务的这 $N$ 个基站的准确统计信息情况下最大化其可以获得的总奖励量。而为了实现这一优化目标,用户需要在尽可能短的时间内分辨出有最高数学期望奖励的那个最优基站。而想要做到这一点就需要对每个基站进行足够的接入尝试,以对其分布做足够的采样估计。在实际系统中,若以最大化奖励为目标,则用户所采用的决策就必须平衡好“探索”与“利用”之间的均衡。一方面,用户为了增加眼前的奖励,决策应偏向尽可能“利用”经验最佳基站;另一方面,如果是从长远角度考虑,用户需要更多地“探索”其他基站以找到可能潜在的性能更佳的基站。为平衡这种关系,一种流行的方法在统计学以及机器学习领域内广泛使用,该方法是将“探索”与“开发”困境建模为多臂老虎机问题[11]。

解决多臂老虎机问题的经典算法为置信上界算法[12]。该方法为多臂老虎机的 $N$ 个操纵杆维护它们各自的置信上界

$$\hat{u}_i + \sqrt{\frac{2\log t}{T_i}}, \tag{8}$$

其中:$\hat{u}_i$ 是对操纵杆 $i$ 的采用均值,$T_i$ 为截止时隙 $t$ 为止操纵杆 $i$ 的采样次数。可以看出第 2 项的作用是增加探索相比之下尚未充分采样的基站的奖励。这 2 项结合在一起可以很好地平衡“探索”与“利用”之间的均衡。

用户在每一个时隙开始时做出决策决定其在该时隙内选择建立链接的基站。我们称用户触发了一次网络切换为用户在相邻的 2 个时隙选择了不同的基站接入。用户触发一次网络切换会造成一些额外的损失,例如,时延、能量消耗或者是额外的频谱资源开销用来发送网络切换的控制信号。所以,频繁的网络切换很可能会严重降低服务质量(quality of service, QoS)以及体验质量(quality of experience, QoE)。与文献[8]及文献[10]类似,假设每一次切换会导致固定的损失 $C$,那么在使用给定算法 $\Pi$ 情况下,截止到时隙 $T$ 为止,用户的网络切换损失可以表达为

$$H^{\Pi}(\boldsymbol{\Theta}) = C\sum_{i=1}^{N}\sum_{t=1}^{T} 1_{a_t = i, a_{t-1} \neq i}, \tag{9}$$

这里 $1_S$ 表示事件 $S$ 的指示函数。

由此可见,虽然多臂老虎机模型很好地刻画了“探索”与“利用”之间的均衡问题,但是这种方法无法直接运用于用户接入问题,因为该算法需要在各个基站之间反复切换。而该问题在无线网络中会尤其严重。考虑如下场景,当拥有最高数学期望奖励的 2 个基站之间的数学期望差距 $\Delta_{N-1}$ 非常小的时候,该算法很难区分哪一个才是最好的,从而陷入在二者之间反复切换的僵局。为克服这一问题,在文献[13]中,算法设置在运行足够长的时间后终止探索过程,进而选择任意一个足够好的基站一直建立链接。

## 1.2　在线学习中的后悔值与切换次数上界

为评价以置信上界算法等为代表的在线学习算法的性能优劣,常用的性能指标为后悔度。该指标定义为截止到 $T$ 时所获得的总奖励与一直采用最高奖励均值的决策之间的差值。通常在理论分析时,关注的是后悔度的数学期望

$$R^{\Pi}(T) = E\left[T \cdot u_N - \sum_{t=1}^{T} u(t, a_t)\right]. \tag{10}$$

Auer 等[12]的早期工作已经表明传统置信上界算法的期望后悔度有严格上界:$O(\log T)$,这意味着该算法的后悔值是时隙 $T$ 的高阶无穷小

量,但同时文献也指出该算法运行时在不同操作杆之间切换次数的数学期望上界也是 $O(\log T)$。而这就意味着在无线密集网络问题中,由于网络切换带来的损失不可忽略,置信上界算法无法直接运用。

Shen 和 Schaar[9] 提出一种直接的改进置信上界算法并给出了理论分析。在该算法中,连续的 $k$ 个时隙会被组合在一起成为一个大时隙,并且 $k$ 会从 1 开始逐一增加。在这连续的 $k$ 个时隙内,用户只在第一个时隙开始时做出决策选择基站并一直与该基站保持接入。作者证明了在该策略下,其依旧能保证以 $O(\log T)$ 为上界的期望后悔度,但用户在不同基站之间切换次数的数学期望上界可以降低为: $o(\log T)$。这样一来就可以保证用户在不同基站之间切换造成的损失在阶数上可以忽略不计。

## 2 高效的用户接入策略

本节首先提出基于操作杆淘汰机制的一种用户接入算法。随后的理论分析可以表明,该算法在保持期望后悔度上界 $O(\log T)$ 不变的情况下,可以将用户在不同基站之间切换次数的数学期望上界降低为常数阶。相比较于前文提到的算法有重要改进。

### 2.1 基于淘汰机制的用户接入算法

本节提出的算法以回合的形式来运行。每个回合包含若干连续个时隙。在算法运行过程中需要引入一个监控变量 $\hat{\Delta}_m$ 用来估计与监控所有的次优基站和最优基站之间的差距,其中下标 $m$ 表示第 $m$ 回合。$\hat{\Delta}_m$ 在每回合结束时会减半,最终得到 $\hat{\Delta}_m < \Delta_i, \forall i \in [1, N-1]$。每个回合都包含 3 个步骤:均等采样、参数更新、操作杆淘汰。

每个回合的均等采样步骤所需要的时隙总数由当前所剩余的候选基站总数和当前的回合数所决定。用户在该步骤对每个基站保持接入连续的 $L_m$ 个时隙

$$L_m = n_m - n_{m-1}, \tag{11}$$

其中 $n_m$ 的定义如下

$$n_m := \frac{2\log(CT\hat{\Delta}_m^2)}{\hat{\Delta}_m^2}. \tag{12}$$

用户在完成上述步骤之后会用其所获得的奖励数据来更新参数,包括均值的更新、监控变量的更新和采样次数的更新。

每个回合的最后阶段,用户会根据更新后的系统参数来淘汰掉被分类判断为次优的基站,也就是说在这之后的所有回合用户都不会再选择这个基站接入。这里设置淘汰判决条件为

$$\hat{u}_i + \sqrt{\frac{\log(CT\hat{\Delta}_m^2)}{2n_m}} < \max_{j \in B_m}\left\{\hat{u}_j - \sqrt{\frac{\log(CT\hat{\Delta}_m^2)}{2n_m}}\right\}, \tag{13}$$

其中 $B_m$ 为当前回合 $m$ 尚未被淘汰的基站集合。在该回合之后所有满足上述判决条件的基站 $i$ 都会被淘汰。

完整算法的具体流程见算法 1。

**算法 1** UAAE(user association with arm-elimination)用户接入算法:

**步骤 1)** 接受输入时间长度 $T$, 切换损失 $C$, 初始化监控变量等 $\hat{\Delta}_0 = b, t = 1, m = 1$, 以及 $B_0 = B$。

**步骤 2)** 若 $|B_m| = 1$, 则用户保持接入 $B_m$ 中的唯一基站直到时隙 $T$, 否则执行步骤 3)。

**步骤 3)** 均等采样:依次选择接入 $|B_m|$ 中的基站,并且每个基站保持接入 $L_m$ 个时隙,计算从每个基站 $i$ 获得的总奖励

$$r_i = \sum_{t'=t}^{t+L_m|B_m|-1} u(t', i), \tag{14}$$

**步骤 4)** 参数更新:按以上得到数据更新如下系统参数

$$\hat{u}_i = \hat{u}_i + \frac{1}{n_m}(r_i - L_m\hat{u}_i), \tag{15}$$

$$t = t + L_m|B_m|, \tag{16}$$

$$\hat{\Delta}_{m+1} = \frac{\hat{\Delta}_m}{2}, \tag{17}$$

**步骤 5)** 操作杆淘汰:按照新的系统参数以及淘汰判决条件(13)更新 $B_m$,

$$m = m + 1, \tag{18}$$

**步骤 6)** 若 $t > T$, 算法结束,否则跳转返回至步骤 2)。其中,$|S|$ 表示集合 $S$ 的势。

### 2.2 理论分析

**定理 1** 算法 1 保持了和置信上界相同的后悔度数学期望上界,同时用户在不同基站之间切换次数的数学期望上界可以降低为常数阶:

a) 算法 1 可以保证如下后悔度的上界

$$R(T) \leqslant \sum_{i=1}^{N-1}\left(\Delta_i + \frac{32\log(CT\Delta_i^2)}{\Delta_i} + \frac{96}{C\Delta_i}\right), \tag{19}$$

b) 对于每一个次优基站 $i$, 设置变量 $m_i$ 为第一次满足如下条件的回合数: $\hat{\Delta}_{m_i} < \Delta_i$, 也就是说

$$m_i = \min\{m \mid \hat{\Delta}_m < \Delta_i\}, \tag{20}$$

算法 1 可以保证如下切换次数的上界

$$E[H(\Theta)] \leqslant \sum_{j=0}^{m_{N-1}} \frac{(N-1)b^2}{C\hat{\Delta}_j^2 \log(Cb^2)} + \sum_{i=1}^{N-1}\left((N-i+1)(m_i - m_{i-1}) + \frac{b^2}{C\hat{\Delta}_{m_i}^2 \log(Cb^2)}\right), \tag{21}$$

**证明**　为描述简洁,首先需要定义 3 个事件: $E_1, E_2, E_3$。

1) $E_1$:每一个次优基站 $i$ 都会在第 $m_i + 1$ 回合之前满足淘汰判决条件,从而被淘汰掉不再被使用;

2) $E_2$:存在某一个次优基站 $i$ 没有在第 $m_i + 1$ 回合之前满足淘汰判决条件,被继续使用;

3) $E_3$:最优基站被错误淘汰。

这里需要注意,以上事件分解得到的 3 个子集的并集是事件全集,但是交集未必是空集。

当事件 $E_1$ 发生时,每一个次优基站 $i$ 都只能提供接入至多 $m_i$ 回合。在所有的次优基站都被淘汰掉以后,网络用户便不会再触发网络切换,由此可以得到在该事件发生的情况下用户触发网络切换次数的上界

$$F_1 = \sum_{i=1}^{N-1}(N-i+1)(m_i - m_{i-1}), \tag{22}$$

其中 $m_0 = 0$。

相应地,由于每一个次优基站 $i$ 被选择接入的次数都不超过 $n_{m_i}$。结合 $\Delta_i$ 并且通过不等式放缩,可以得到当事件 $E_1$ 发生时,所造成的后悔度

$$\sum_{i=1}^{N-1}\Delta_i n_{m_i} = \sum_{i=1}^{N-1}\Delta_i \frac{2\log(CT\hat{\Delta}_m^2)}{\hat{\Delta}_m^2} \leqslant \sum_{i=1}^{N-1}\left(\Delta_i + \frac{32\log(CT\Delta_i^2)}{\Delta_i}\right), \tag{23}$$

当事件 $E_2$ 发生时,由文献[14]可以得知某一个次优基站 $i$ 没有在第 $m_i + 1$ 回合之前满足淘汰判决条件的发生概率有上界: $\frac{2}{CT\hat{\Delta}_{m_i}^2}$, 可以利用如下不等式放缩来得到在该事件发生的情况下用户触发网络切换次数的上界

$$\frac{2}{CT\hat{\Delta}_{m_i}^2} \times \frac{Tb^2}{2\log(CTb^2)} = \frac{b^2}{C\hat{\Delta}_{m_i}^2 \log(CTb^2)} < \frac{b^2}{C\hat{\Delta}_{m_i}^2 \log(Cb^2)}, \tag{24}$$

这里利用用户最大可能触发的网络切换次数

$$D_{\max} = \frac{T}{\frac{2\log(CT\hat{\Delta}_0^2)}{\hat{\Delta}_0^2}} \leqslant \frac{Tb^2}{2\log(CTb^2)}. \tag{25}$$

上式成立的条件也是基于事实在均等采样步骤中每个基站保持接入 $L_m$ 个时隙,而 $L_m$ 是单调递增的。对所有 $N-1$ 个次优基站求和,可以得到针对事件 $E_2$ 的用户触发网络切换次数的上界

$$F_2 = \sum_{i=1}^{N-1}\frac{b^2}{C\hat{\Delta}_{m_i}^2 \log(Cb^2)}. \tag{26}$$

相应地,当事件 $E_2$ 发生时,其造成的后悔度可以如下表达:

$$\sum_{i=1}^{N-1}\Delta_i \frac{2}{CT\hat{\Delta}_{m_i}^2}T \leqslant \sum_{i=1}^{N-1}\frac{8}{C\hat{\Delta}_{m_i}^2} \leqslant \sum_{i=1}^{N-1}\frac{32}{C\Delta_i}. \tag{27}$$

当事件 $E_3$ 发生时,同样可以由文献[14]得知最优基站被某一个次优基站 $i$ 所错误淘汰的概率也有上界 $\frac{2}{C\,T\hat{\Delta}_{m_i}^2}$。再次利用式(24)中的最大次数 $D_{\max}$, 可以得到针对事件 $E_3$ 的用户触发网络切换次数的上界

$$F_3 = \sum_{j=0}^{m_{N-1}}(N-1)\frac{2}{CT\hat{\Delta}_j^2}\frac{Tb^2}{2\log(CTb^2)} < \sum_{j=0}^{m_{N-1}}\frac{(N-1)b^2}{C\hat{\Delta}_j^2\log(Cb^2)}, \tag{28}$$

注意该求和上下限为从 0 到 $m_{N-1}$ 回合,这是因为最优基站在 $m_{N-1}$ 回合之后被淘汰的情况实际上已经被包含在了事件 $E_2$ 当中。

相应地,当事件 $E_3$ 发生时,记最优基站在 $m_*$ 回合被淘汰,那么最优基站被淘汰后的每一个时隙都会增加后悔值。这里可以将该情况下的后悔值做如下不等式放缩:

$$\sum_{m_*=0}^{m_{N-1}}\sum_{i,m_i\geqslant m_*}\frac{2}{CT\hat{\Delta}_{m_*}^2}T\max_{j,m_j\geqslant m_*}\Delta_j \leqslant \sum_{m_*=0}^{m_{N-1}}\sum_{i,m_i\geqslant m_*}\frac{2}{C\hat{\Delta}_{m_*}^2}4\hat{\Delta}_{m_*} = \sum_{i=1}^{N-1}\sum_{m_*=0}^{m_i}\frac{8}{C2^{-m_*}} \leqslant \sum_{i=1}^{N-1}\frac{64}{C\Delta_i}. \tag{29}$$

结合以上事件分解后的分部结论相加,可以得到定理 1 中的结论。

证明完毕。

在该算法中,用户对每个基站保持接入连续的 $L_m$ 个时隙,而 $L_m$ 可以被验证是指数增长的。文献[9]中的算法是逐一增长的。这样做的好处是更有利于减缓前文中已经提到的当拥有最高数学期望奖励的 2 个基站之间的数学期望差距 $\Delta_{N-1}$ 非常小的时候,算法会很难区分这两者的问题。而从另一方面来说,非常小的 $\Delta_{N-1}$ 也保证了即使用户需要在较长的连续时隙内选择次优的那个基站接入,也不会造成特别大的后悔度。

$L_m$ 同时也会随着 $T$ 的增加而增加。这个设计能够保证如果总时间足够长,算法会在每个回合执行淘汰机制之前对当前回合还存在的基站做充分的估计。同时由于总时间足够长,在所有次优基站被淘汰前对它们进行采样所使用的时隙造成的后悔度也相对影响较小。

该算法所保障的低切换的特性为更实际的网络模型提供了很好的使用条件。在实际的网络环境中,网络的各方面性能往往是动态变化的,即前文中所假设的奖励值由一个稳定分布所产生的条件不再成立。而算法往往需要在环境发生变化时重新开始学习以适应新的网络环境。这样一来,如果使用常规的多臂老虎机算法,则不可避免地重新造成了大量的网络切换。所以,在动态环境下,当用户需要重新开始学习过程时,UAAE 算法能够保证节省大量的切换次数。

## 3 实验结果分析

本节将利用蒙特卡洛仿真的方法验证本文所提算法的有效性。所有结果均为 500 次仿真实验的平均结果。

参数设定如下,系统中总共包含 30 个基站。每一个基站的期望奖励值由正态分布产生,$N(5, \delta^2)$。基站 $i$ 在每次被用户选中接入时产生的奖励同样服从正态分布 $N(u_i, 1)$。这样设定的好处是可以通过控制方差 $\delta$ 来调节不同基站之间的优劣差距。$\delta$ 越小意味着 $\triangle_i$ 也越小。仿真的总时间长度为 $T = 5 \times 10^5$,监控变量初始值为 $b = 1$,用户每次网络切换的损失为 $C = 1$。

为对比本文提出算法的有效性,本文选择如下算法作为对比算法:

a) 置信上界算法;

b) 改进置信上界算法[9];

c) $\epsilon$ 贪心算法[11]。该算法在每个时隙都以 $\epsilon$ 概率选择随机接入一个基站,以 $1-\epsilon$ 概率选择接入当前经验最佳,即估计均值最高的基站。设定 $\epsilon = 0.01$。

首先验证 4 种不同算法的后悔值性能表现。如图 2 所示,可以清晰地看出,除 $\epsilon$ 贪心算法的后悔值曲线几乎是线性增长,其他 3 个算法增长的阶数是一致的,但是 UAAE 算法有一定系数上的损失。

图 3 表现的是 4 种算法下用户所触发的网络切换次数随时隙的变化曲线对比。UAAE 算法相较于其他 3 个算法有着绝对的优势。这其中的原因正如前文所解释,用户最终淘汰掉 $N-1$ 个基站,从而不会再触发任何网络切换。$\epsilon$ 贪心算法的网络切换次数为线性增长。另外 2 种算法均为对数增长。

接下来验证不同基站之间的期望奖励差距大小对算法性能的影响及敏感度分析。针对这点的

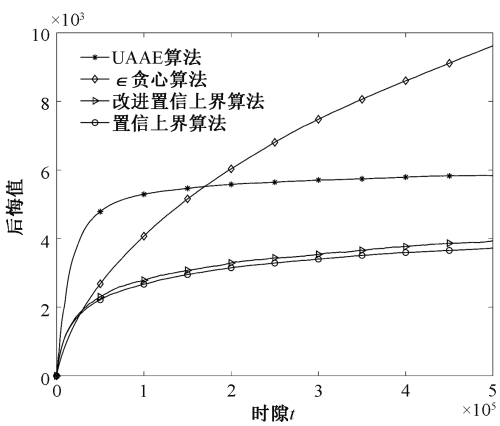


**图 2 不同算法下的后悔值性能指标对比**

Fig. 2 Cumulative regrets with various bandit handover algorithms

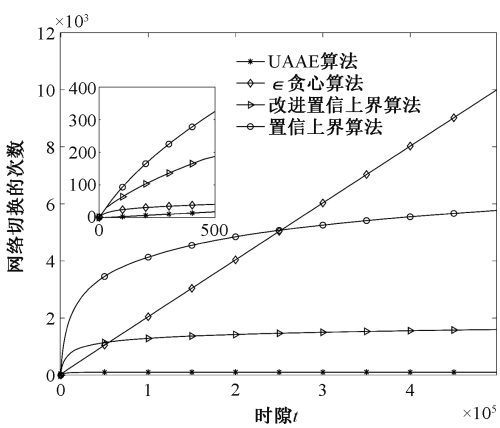


**图 3 不同算法下的用户触发网络切换次数对比**

Fig. 3 Cumulative numbers of handovers with various bandit handover algorithms

研究是必要的,但是往往被很多使用多臂老虎机模型的研究工作所忽略。不同于其他可以使用多臂老虎机的应用系统(例如推荐系统等)。在实际无线网络系统中,用户所能选择接入的基站往往可能不会有太大的性能差距。这就给很多多臂老虎机算法带来了挑战。在该实验中,参数 $\delta$ 变化范围 0.05~0.5。从图 4 可以看出,$\epsilon$ 贪心算法虽然不受参数 $\delta$ 的任何影响,但是用户所触发的网络切换次数也是最多的。另外 3 个算法中,文献[9]中的改进置信上界算法性能要略好于置信上界算法。而本文所提算法 UAAE 可以在任何参数 $\delta$ 下保持最小的网络切换次数。

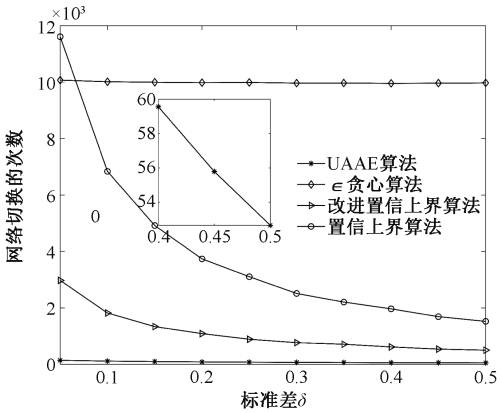


**图 4　不同算法对于基站间差距大小的敏感度对比**

Fig. 4　Sensitivity to the gaps between the SBSs

## 4　结论与讨论

本文利用多臂老虎机模型提出一个低复杂度的无线网络用户接入算法。通过 3 组不同角度的对比试验,验证了本文所提算法的有效性、鲁棒性,为下一代无线通信网络中用户接入系统设计提供一种解决思路。该算法除了有效降低用户触发网络切换的次数,也保证其后悔值性能不会受到影响。

本文主要考虑的是稳定环境下的解决方案,即基站产生奖励值的概率分布是恒定不变的。但是正如前文的讨论中所指出,当用户面临动态变化的网络环境时,其需要频繁重新开始学习过程。而本文所提算法为在动态环境下的部署使用提供了很好的基础。

### 参考文献

[1] Boccardi F, Heath R W, Lozano A, et al. Five disruptive technology directions for 5G[J]. IEEE Communications Magazine, 1996, 52(2): 74-80. DOI: 10. 1109/MCOM. 2014. 6736746.

[2] Lee W, Cho D H. Enhanced group handover scheme in multiaccess networks[J]. IEEE Transactions on Vehicular Technology, 2011, 60(5): 2389-2395. DOI:10. 1109/TVT. 2011. 2140386.

[3] Fischione C, Athanasiou G, Santucci F. Dynamic optimization of generalized least squares handover algorithms [J]. IEEE Transactions on Wireless Communications, 2014, 13(3): 1235-1249. DOI: 10. 1109/TWC. 2014. 013014. 121720.

[4] Guidolin F, Pappalardo I, Zanella A, et al. Context-aware handover policies in HetNets[J]. IEEE Transactions on Wireless Communications, 2016, 15(3): 1895-1906. DOI: 10. 1109/TWC. 2015. 2496958.

[5] Ye Q Y, Rong B Y, Chen Y D, et al. User association for load balancing in heterogeneous cellular networks[J]. IEEE Transactions on Wireless Communications, 2013, 12(6): 2706-2716. DOI:10. 1109/TWC. 2013. 040413. 120676.

[6] Videv S, Haas H. Energy-efficient scheduling and bandwidth-energy efficiency trade-off with low load[C]//2011 IEEE International Conference on Communications(ICC). June 5-9, 2011, Kyoto, Japan. IEEE, 2011: 1-5. DOI:10. 1109/icc. 2011. 5962571.

[7] 孟庆民,赵媛媛,岳文静,等. 动态超密集网络中的 Markov 预测切换[J]. 通信学报, 2018, 39(10): 166-174. DOI: 10. 11959/j. issn. 1000-436x. 2018225.

[8] Wang Z, Li L H, Xu Y, et al. Handover control in wireless systems via asynchronous multiuser deep reinforcement learning[J]. IEEE Internet of Things Journal, 2018, 5(6): 4296-4307. DOI:10. 1109/jiot. 2018. 2848295.

[9] Shen C, van der Schaar M. A learning approach to frequent handover mitigations in 3GPP mobility protocols[C]//2017 IEEE Wireless Communications and Networking Conference. March 19-22, 2017, San Francisco, CA, USA. IEEE, 2017: 1-6. DOI:10. 1109/WCNC. 2017. 7925950.

[10] Zhou Y M, Shen C, van der Schaar M. A non-stationary online learning approach to mobility management[J]. IEEE Transactions on Wireless Communications, 2019, 18(2): 1434-1446. DOI:10. 1109/TWC. 2019. 2893168.

[11] Bubeck S. Regret analysis of stochastic and nonstochastic multi-armed bandit problems[M]. Boston: Now Publishers Inc, 2012. DOI:10. 1561/9781601986276.

[12] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem[J]. Machine Learning, 2002,47: 235-256. DOI:10. 1023/A:1013689704352.

[13] Sun Y X, Zhou S, Xu J. EMM: energy-aware mobility management for mobile edge computing in ultra dense networks[J]. IEEE Journal on Selected Areas in Communications, 2017, 35(11): 2637-2646. DOI: 10. 1109/JSAC. 2017. 2760160.

[14] Auer P, Ortner R. UCB revisited: improved regret bounds for the stochastic multi-armed bandit problem[J]. Periodica Mathematica Hungarica, 2010, 61(1/2): 55-65. DOI:10. 1007/S10998-010-3055. 6.