Classification Bandits: Classification Using Expected Rewards as Imperfect Discriminators

Koji Tabata[1,3(✉)], Atsuyoshi Nakumura[2], and Tamiki Komatsuzaki[1,3]

[1] Institute for Chemical Reaction Design and Discovery, Hokkaido University, Sapporo, Japan
ktabata@es.hokudai.ac.jp
[2] Graduate School/Faculty of Information Science and Technology, Hokkaido University, Sapporo, Japan
[3] Research Center of Mathematics for Social Creativity Research Institute for Electronic Science, Hokkaido University, Sapporo, Japan

Abstract. A classification bandits problem is a new class of multi-armed bandits problems in which an agent must classify a given set of arms into positive or negative depending on whether the number of bad arms are at least N_2 or at most $N_1 (< N_2)$ by drawing as fewer arms as possible. In our problem setting, bad arms are imperfectly characterized as the arms with above-threshold expected rewards (losses). We develop a method of reducing classification bandits to simpler one threshold classification bandits and propose an algorithm for the problem that classifies a given set of arms correctly with a specified confidence. Our numerical experiments demonstrate effectiveness of our proposed method.

Keywords: Multi-armed bandits · Threshold bandits · Thompson sampling

1 Introduction

How to determine the presence or extent of a disease from biopsy under the existence of some uncertainty is of crucial importance in life science. Let us consider the following cancer diagnosis problem in which a doctor has to diagnose whether a certain patient has cancer or not from given his/her K cells: if the number of cancer cells N is negligible ($N \leq N_1$), then the doctor can diagnose that the patient does not have cancer, but if it is non-negligible ($N \geq N_2 > N_1$), the doctor should diagnose that the patient does. One of the cancer cell diagnosis methods is the classification of cells in terms of a set of Raman spectra[1] averaged over each cell [8]. However, Raman measurements require more than ten hours

[1] Histopathologists usually diagnose whether cells are of cancer or not by inspecting their morphological characteristics with a human bias, but Raman measurements are considered to enable more reliably to judge the cell states.

M. Gupta and G. Ramakrishnan (Eds.): PAKDD 2021 Workshops, LNAI 12705, pp. 57–69, 2021.
https://doi.org/10.1007/978-3-030-75015-2_6

for one hundred cells, by scanning point illumination to single cells along which Raman spectra are acquired in time to time. Thus, interactive measurement depending on Raman spectra obtained so far is a key to realize fast cell diagnosis. Such interactive measurement can be formulated as a bandit problem treated in this paper by regarding each cell as an arm, and letting Raman spectra sampling from each cell correspond to an arm draw.

In this bandit problem, doctor cannot always conclude whether the number of cancer cells is negligible or not correctly due to two different types of uncertainty. The first type of uncertainty is the variance of reward (cancer index calculated from sampled Raman spectra) obtained by each arm draw, which has been extensively studied in the area of statistics. The second type of uncertainty is the imprecision (imperfect positive predictive value) of the true expected reward of each arm (cancer index averaged over each cell) and this frequently happens in real situation. In fact, Raman spectra averaged over each cell was reported to be classified into cancer or normal cells with about 85% accuracy [8]. While we can obtain as much accurate value as we want by taking enough number of samples for the first type of uncertainty, the second type of uncertainty cannot be reduced. To the best of our knowledge, this paper is the first one taking account of the second type of uncertainty in the context of bandits.

In this paper, we study a pure exploration K-armed bandit problem, named *classification bandit problem*, in which an agent must classify a given set of K arms into "negative" or "positive" depending on whether the number of bad arms is at most N_1 or at least $N_2 (> N_1)$, respectively, within given allowable failure probabilities δ_N and δ_P by drawing as small number of arms as possible. In our formulation, the mean reward μ_i of each arm i is assumed to be an imperfect discriminator of badness; $\mu_i \geq \theta$ holds for each bad arm i with probability $p_{\mathrm{TP}} > 0.5$ and $\mu_i < \theta$ holds for each good arm i with probability $p_{\mathrm{TN}} > 0.5$.

We show that the classification bandit problem with parameters $p_{\mathrm{TP}}, p_{\mathrm{TN}}, \delta_N, \delta_P, N_1, N_2$ can be reduced to the problem, named *one-threshold classification bandit problem*, which has only one threshold λ instead of two thresholds N_1 and N_2 and one allowable failure probability δ instead of δ_P and δ_N and is free from the second type of uncertainty. Our reduction is not always possible, and we show the condition of the reduction and how to calculate λ and δ from $p_{\mathrm{TP}}, p_{\mathrm{TN}}, \delta_N, \delta_P, N_1, N_2$. For the one-threshold classification bandit problem, we propose an algorithm and prove its correctness for any arm selection policy. We also propose a Thompson-sampling-based arm selection policy for this algorithm, which demonstrates a faster stopping time of the algorithm than UCB-based and Successive-Elimination-based arm selection policies.

Related Works

One-threshold classification bandits problem is regarded as a kind of pure exploration bandit problem. Complexity analysis of this type of problem is performed by minimizing error probability under a fixed budget or minimizing the number

of samples under a fixed confidence. In this paper we focus on the fixed confidence setting. The most studied pure exploration bandit problem is best arm identification whose objective is to find the k highest expected reward arms for a given k. For the best arm identification problem, Audibert et al. [1] proposed an algorithm based on successive elimination that eliminates the worst arm one by one from candidates of the best arm. Later, more efficient algorithms that are not based on elimination such as LUCB [3] and UGapE [2] were proposed.

Instead of the highest expected reward arms, Locatelli et al. [6] proposed an algorithm for a thresholding bandit problem in which an agent has to output all the arms with the expected reward higher than a given threshold. Kano et al. [4] formulated the good arm identification problem whose task is to find λ arms whose expected rewards are above a given threshold, for a given λ, if at least such λ arms exist, or to find all such arms otherwise. Kaufmann et al. [5] and Tabata et al. [10] independently studied a problem to decide whether at least one arm exists or not, whose expected reward is above the threshold, in which precise identification of such arms is not necessarily required.

A question of 'how one can derive accurate decision under the condition that only qualitative test results would be obtained' is one of the most intriguing subjects in the area of fault detection of systems. In many cases, several kinds of tests are assumed to be given explicitly with their false-positive, and the false-negative rates. Nachla et al. [7] treated a problem to design the permutation of tests in order to decrease the total cost of, e.g., quality inspections, repairs of good components in products, and dispositions of no-fault systems, and proposed heuristics to solve the problem. Raghavan et al. [9] proposed a method to decide an optimal test sequence with qualitative tests by using dynamic programming.

To our best knowledge, there exists no research that deals with the problem of accurate decision in terms of qualitative test results in the context of bandits algorithm. In this paper, we present an algorithm to transform classification bandits based on qualitative tests with nonnegligible false-positive and false-negative rates into one-threshold bandits problem which can address the transformability to one threshold bandits and design the threshold and the error rate to meet the given allowed error rates.

2 Problem Formulation

We study a variant of a K-armed bandit problem defined as follows. An agent is given a set of K arms that is composed of *bad* and *good* arms. No information about which arm is bad or good is directly provided to the agent. However, the agent can get values $X_i(1), X_i(2), \cdots$ of an indicator to represent badness for each arm i by drawing it repeatedly, where $X_i(n)$ denotes the value obtained by the nth draw of arm i. For each arm i, we assume that $X_i(1), X_i(2), \cdots$ are i.i.d. random variables whose distribution is denoted as ν_i with mean μ_i (whose distribution and value cannot be known *a priori*).

For a given threshold θ, an arm i satisfying $\mu_i \geq \theta$ ($\mu_i < \theta$) is defined as *positive (negative) arm*. We consider the case that arm's expected indicator value μ_i is an imperfect discriminator of its badness or goodness; a bad

arm is positive with probability p_{TP} and a good arm is negative with probability p_{TN}. We also assume that, whether each arm is positive or negative, is independent of any different arms i and j, e.g., $P[\text{arms } i \text{ and } j \text{ are positive}] = P[\text{arm } i \text{ is positive}]P[\text{arm } j \text{ is positive}]$.

At every time step $t = 1, 2, \ldots$, an agent chooses one of the K arms $i_t \in \{1, 2, \ldots, K\}$ and gets an indicator value $X_{i_t}(n_{i_t}(t))$ of badness that is drawn from distribution ν_{i_t}, where $n_i(t)$ is the number of times arm i has been drawn by time step t. The agent's task is to conclude whether the given set of K arms contains non-negligible number of bad arms (i.e., a number of bad arms enough to judge "positive") or not by drawing arms as few times as possible. We can formulate the problem as follows.

Problem 1 (classification bandits). For given $p_{\mathrm{TP}}, p_{\mathrm{TN}} \in (0.5, 1]$ and $\delta_P, \delta_N \in (0, 0.5)$, output "negative" with probability at least $1 - \delta_N$ if the number of bad arms is less than or equal to N_1, and output "positive" with probability at least $1 - \delta_P$ if the number of bad arms is larger than or equal to N_2 by drawing as small number of arms as possible.

Note that there is no requirement for the output when the number of bad arms is larger than N_1 and lower than N_2.

3 Problem Reduction

3.1 Reduction Theorem

In this section, we show how to reduce classification bandits, which contain uncertainty derived from probabilities p_{TP} and p_{TN}, to the following one-threshold classification bandits, which is free from such uncertainty.

Problem 2 (one-threshold classification bandits). For a given $\lambda \in \mathbb{N}$ and $\delta > 0$, output "negative" with probability at least $1 - \delta$ if the number of positive arms is less than λ, and output "positive" with probability at least $1 - \delta$ if the number of positive arms is at least λ by drawing as small number of arms as possible.

In this problem setting, the number to be identified is not the number of bad arms but that of positive arms, that is, we only consider the uncertainty due to reward variance to solve this problem. We will introduce the algorithm to solve this reduced problem in the next section. Here we explain how such reduction is possible.

Given the number of arms K, the number of bad arms $N \in [0, K]$, probability $p_{\mathrm{TP}} > 0.5$ with which a bad arm is positive, and probability $p_{\mathrm{TN}} > 0.5$ with which a good arm is negative, the probability-generating function $G^{(N,K)}(t)$ for the number of positive arms is expressed as

$$G^{(N,K)}(t) = (p_{\mathrm{TP}}t + (1 - p_{\mathrm{TP}}))^N((1 - p_{\mathrm{TN}})t + p_{\mathrm{TN}})^{K-N},$$

because of the arm's independency.

Let $c_d^{(N,K)}$ denote the coefficient of t^d in $G^{(N,K)}(t)$. Then, $c_d^{(N,K)}$ is the probability that the number of positive arms is d in the case that the given set of K arms contains just N bad arms.

The following proposition is used to prove Lemma 1.

Proposition 1. *$\sum_{j=0}^{\ell-1} c_j^{(N,K)}$ is a weakly decreasing function on N and $\sum_{j=\ell}^{K} c_j^{(N,K)}$ is a weakly increasing function on N.*

Proof. Omitted due to space limitations. □

Let X denote the number of positive arms. Then, $P[X \geq \ell | N = i]$ and $P[X < \ell | N = i]$, probabilities of $X \geq \ell$ and $X < \ell$ under the condition of $N = i$, are $\sum_{j=\ell}^{K} c_j^{(i,K)}$ and $\sum_{j=0}^{\ell-1} c_j^{(i,K)}$, respectively.

From Proposition 1, we can have the following lemma.

Lemma 1. *Let X and N be the number of positive arms and the number of bad arms, respectively. Then, the following two inequalities hold:*

$$P[X \geq \ell | N \leq N_1] \leq P[X \geq \ell | N = N_1] \text{ and}$$
$$P[X < \ell | N \geq N_2] \leq P[X < \ell | N = N_2].$$

Proof. By Proposition 1, for $i \leq N_1$,

$$P[X \geq \ell | N = i] = \sum_{j=\ell}^{K} c_j^{(i,K)} \leq \sum_{j=\ell}^{K} c_j^{(N_1,K)} = P[X \geq \ell | N = N_1]$$

holds, and thus

$$\begin{aligned} P[X \geq \ell | N \leq N_1] &= \frac{\sum_{i=0}^{N_1} P[X \geq \ell | N = i] P[N = i]}{\sum_{i=0}^{N_1} P[N = i]} \\ &\leq \frac{\sum_{i=0}^{N_1} P[X \geq \ell | N = N_1] P[N = i]}{\sum_{i=0}^{N_1} P[N = i]} = P[X \geq \ell | N = N_1]. \end{aligned}$$

We can show the second inequality similarly. □

This implies that, e.g., the failure probability $P[X \geq \ell | N \leq N_1]$, that is, under the condition of $N \leq N_1$ to be identified as negative, the classifier answers "positive ($X \geq \ell$)", can be upperbounded by $P[X \geq \ell | N = N_1]$. By the above lemma, the following reduction theorem can be proved.

Theorem 1 (reduction theorem). *For classification bandit problem with $p_{TN}, p_{TP} \in (0.5, 1]$, $\delta_P, \delta_N \in (0, 0.5)$, $0 \leq N_1 < N_2 \leq K$, consider the one-threshold classification bandit problem with*

$$\lambda = \arg \max_{\ell} \delta(\ell, p_{TN}, p_{TP}, N_1, N_2, K), \tag{1}$$

$$\delta = \delta(\lambda, p_{TN}, p_{TP}, N_1, N_2, K), \tag{2}$$

where $\delta(\ell, p_{TN}, p_{TP}, N_1, N_2, K) = \min\left(\delta_N - \sum_{j=\ell}^{K} c_j^{(N_1,K)}, \delta_P - \sum_{j=0}^{\ell-1} c_j^{(N_2,K)}\right)$.

In the case with $\delta > 0$*, classification bandit problem associated with nonzero false-positive and false-negative rates can be reduced to a one-threshold bandit problem.*

Proof. Assume that $\delta > 0$ and algorithm A is an algorithm for one-threshold bandit problem with λ and δ defined Eqs. (1) and (2). Consider the case that the number of bad arms N is at most N_1, which is the case that the algorithm is desired to output "negative." Let X be the number of positive arms. The failure probability that algorithm A outputs "positive" falsely is upper-bounded by $P[X \geq \lambda | N \leq N_1] + \delta$. By Eq. (2) and Lemma 1, $\delta \leq \delta_N - P[X \geq \lambda | N = N_1] \leq \delta_N - P[X \geq \lambda | N \leq N_1]$ holds. Thus, $P[X \geq \lambda | N \leq N_1] + \delta \leq \delta_N$ holds. For the case that the number of bad arms N is at least N_2, algorithm A can be proved similarly to output "negative" with probability at most δ_P. □

3.2 Reducible Parameter Region

In the one-threshold classification bandits problem that is reduced from classification bandits problem, the value of confidence parameter δ calculated by reduction theorem, that is, $\delta = \max_\ell \min\left(\delta_N - \sum_{j=\ell}^{K} c_j^{(N_1,K)}, \delta_P - \sum_{j=0}^{\ell-1} c_j^{(N_2,K)}\right)$, significantly affects sample complexity of the problem, and the problem is not solvable if this δ is non-positive. In the followings, we derive an approximate boundary $\delta = 0$ between solvable and unsolvable (N_1, N_2)-region for fixed $p_{\mathrm{TP}}, p_{\mathrm{TN}}, \delta_P, \delta_N$.

Let $X(m)$ be the number of positive arms when the number of bad arms is m. Then, $X(m)$ can be expressed as $X(m) = X_B + X_G$ using $X_B \sim B(m, p_{\mathrm{TP}})$ (B:binomial distribution) and $X_G \sim B(K-m, 1-p_{\mathrm{TN}})$. By law of large numbers, $B(m, p_{\mathrm{TP}}) \approx N(mp_{\mathrm{TP}}, mp_{\mathrm{TP}}(1-p_{\mathrm{TP}}))$ and $B(K-m, 1-p_{\mathrm{TN}}) \approx N((K-m)(1-p_{\mathrm{TN}}), (K-m)p_{\mathrm{TN}}(1-p_{\mathrm{TN}}))$ when $mp_{\mathrm{TP}}, m(1-p_{\mathrm{TP}}), (K-m)(1-p_{\mathrm{TN}})$ and $(K-m)p_{\mathrm{TN}}$ are enough large (e.g. ≥ 5). By reproductive property of normal distribution, $X(m) \sim N(\mu_m, \sigma_m^2)$ holds for $\mu_m = mp_{\mathrm{TP}} + (K-m)(1-p_{\mathrm{TN}})$ and $\sigma_m^2 = mp_{\mathrm{TP}}(1-p_{\mathrm{TP}}) + (K-m)p_{\mathrm{TN}}(1-p_{\mathrm{TN}})$.

By the Polya's approximation $\frac{1}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{t^2}{2}\right) dt \approx \frac{1}{2}\sqrt{1 - \exp\left(-\frac{2}{\pi}x^2\right)}$, $\mathbb{P}\left[X(m) \geq \mu_m + \alpha_\delta \sigma_m\right] \approx \delta$ and $\mathbb{P}\left[X(m) \leq \mu_m - \alpha_\delta \sigma_m\right] \approx \delta$ hold for $\alpha_\delta = \sqrt{\frac{\pi}{2} \ln \frac{1}{1-(1-2\delta)^2}}$. Thus, $\mathbb{P}\left[X(N_1) \geq \mu_{N_1} + \alpha_{\delta_N} \sigma_{N_1}\right] \approx \delta_N$ and $\mathbb{P}\left[X(N_2) \leq \mu_{N_2} - \alpha_{\delta_P} \sigma_{N_2}\right] \approx \delta_P$ hold. Therefore,

$$\begin{aligned} \delta > 0 \Leftrightarrow & \exists \lambda \in \mathbb{N} \text{ s.t. } \mu_{N_1} + \alpha_{\delta_N} \sigma_{N_1} < \lambda < \mu_{N_2} - \alpha_{\delta_P} \sigma_{N_2} \\ \Leftrightarrow & \mu_{N_1} + \alpha_{\delta_N} \sigma_{N_1} < \mu_{N_2} - \alpha_{\delta_P} \sigma_{N_2} \end{aligned}$$

holds approximately. By solving $\mu_{N_1} + \alpha_{\delta_N}\sigma_{N_1} \approx \mu_{N_2} - \alpha_{\delta_P}\sigma_{N_2}$ we have the following approximate boundary $\delta \approx 0$ over N_1-N_2 plane:

$$(N_2 - N_1)(p_{\mathrm{TP}} + p_{\mathrm{TN}} - 1) \approx \alpha_{\delta_N}\sqrt{N_1 p_{\mathrm{TP}}(1-p_{\mathrm{TP}}) + (K-N_1)p_{\mathrm{TN}}(1-p_{\mathrm{TN}})} + \alpha_{\delta_P}\sqrt{N_2 p_{\mathrm{TP}}(1-p_{\mathrm{TP}}) + (K-N_2)p_{\mathrm{TN}}(1-p_{\mathrm{TN}})}. \tag{3}$$

In the case with $p_{\mathrm{TN}} = p_{\mathrm{TP}}$, the approximate boundary becomes a simple line:

$$N_2 \approx N_1 + \frac{(\alpha_{\delta_N} + \alpha_{\delta_P})\sqrt{K p_{\mathrm{TP}}(1-p_{\mathrm{TP}})}}{2p_{\mathrm{TP}} - 1} \tag{4}$$

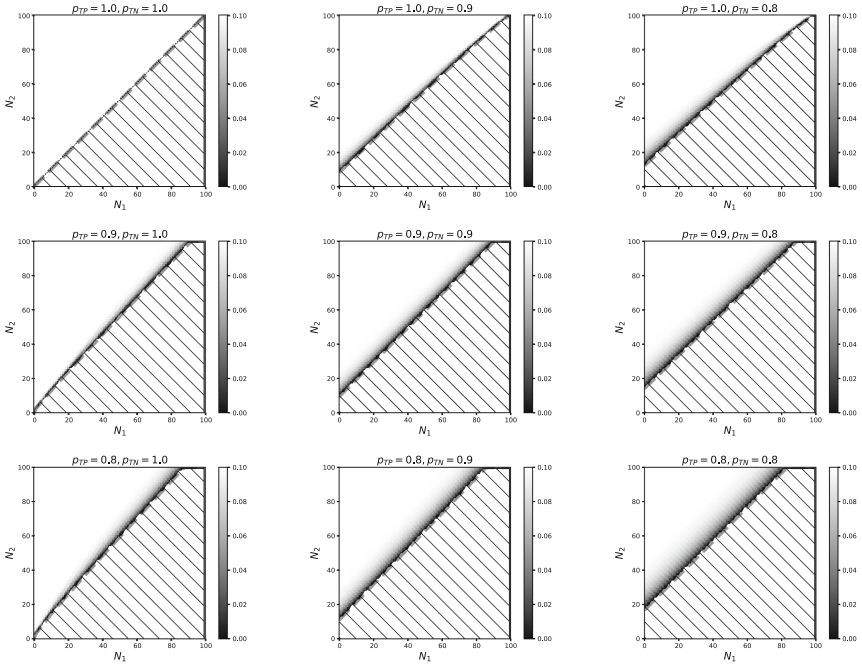

Fig. 1. δ-values at $(N_1, N_2) \in [0, K] \times [0, K]$ for $K = 100$ and $\delta_P = \delta_N = 0.1$ and $(p_{\mathrm{TP}}, p_{\mathrm{TN}}) \in \{1.0, 0.9, 0.8\} \times \{1.0, 0.9, 0.8\}$. $p_{\mathrm{TP}} = 1.0, 0.9, 0.8$ for top, middle and bottom graphs, respectively, and $p_{\mathrm{TN}} = 1.0, 0.9, 0.8$ for left, center and right graphs, respectively. Regions of $\delta < 0$ are filled with oblique lines. The regions of $\delta \geq 0$ are colored in grayscale from black ($\delta = 0$) to white ($\delta = 0.1$). The range of δ is $[-0.9, 0.1]$ because $\delta_P = \delta_N = 0.1$ in these experiments. The approximate boundary by the expression (3) is shown by a gray dashed line on each graph and it looks good approximations.

We give a graphical representation of δ-value over N_1-N_2 plane for fixed K, δ_P, δ_N and various p_{TP}, p_{TN} in Fig. 1.

noend 1. Algorithm for One-Threshold Classification Bandit Problem

Input: K: number of arms, θ: reward threshold, δ: confidence parameter, λ: threshold on the number of positive arms
Output: "positive" or "negative"

initialization: $t \leftarrow 0$, $A \leftarrow \{1, 2, \ldots, K\}$, $D \leftarrow [\,]$, $n_P, n_N \leftarrow 0$, $n_1, \ldots, n_k \leftarrow 0$
loop
  $t \leftarrow t + 1$
  $i_t \leftarrow \mathrm{ASP}(D, K, \delta, \lambda, A)$ {ASP: Arm Selection Policy. See Sec. 4.2}
  Pull arm i_t and update n_{i_t} as $n_{i_t} \leftarrow n_{i_t} + 1$
  Get reward $X_{i_t}(n_{i_t})$ and append $(i_t, X_{i_t}(n_{i_t}))$ to D
  Update $\overline{\mu}_{i_t}(t)$ and $\underline{\mu}_{i_t}(t)$ using Eqs. (5) and (6)
  if $\overline{\mu}_{i_t}(t) < \theta$ **then**
    $A \leftarrow A \setminus \{i_t\}$, $n_N \leftarrow n_N + 1$
    if $K - n_N < \lambda$ **return** "negative"
  else if $\underline{\mu}_{i_t}(t) \geq \theta$ **then**
    $A \leftarrow A \setminus \{i_t\}$, $n_P \leftarrow n_P + 1$
    if $n_P \geq \lambda$ **return** "positive"

For fixed $K = 100$ and $\delta_P = \delta_N = 0.1$, δ-values at $(N_1, N_2) \in [0, K] \times [0, K]$ are shown in Fig. 1 for $(p_{\mathrm{TP}}, p_{\mathrm{TN}}) \in \{1.0, 0.9, 0.8\} \times \{1.0, 0.9, 0.8\}$. The regions of negative δ-values are filled with oblique lines. The regions of non-negative δ-values are colored in grayscale from black ($\delta = 0$) to white ($\delta = 0.1$). From the definition of N_1 and N_2, the region of $N_2 \leq N_1$ is always filled with oblique lines.

When both of p_{TP} and p_{TN} are 1.0, δ is 0.1 at any point of region $N_1 < N_2$, because bad (good) arm is always assigned to positive (negative) arm. Even if there is no bad arm, $(1 - p_{\mathrm{TN}})K$ arms are discriminated as positive arms on average. In fact, at $N_1 = 0$, we have $\delta \leq 0$ for N_2 in some interval $[0, N_2^0]$, and N_2^0 increases as p_{TN} decreases. Similarly, even if all the arms are bad arms, only $p_{\mathrm{TP}}K$ arms are discriminated as positive arms on average. In fact, at $N_2 = K$, we have $\delta \leq 0$ for N_1 in some interval $[N_1^K, K]$, and N_1^K decreases as p_{TP} decreases.

We can see that the expression (3) gives good approximation from the approximate boundaries shown by a gray dashed lines.

4 Algorithm

The pseudocode of proposed algorithm for one-threshold classification bandits is shown in Algorithm 1. Note that the algorithm works for any arm selection policy.

4.1 Decision Condition

The upper and lower confidence bounds $\overline{\mu}_{i,n}$, $\underline{\mu}_{i,n}$ of μ_i after taking n samples from arm i are defined as follows:

$$\overline{\mu}_{i,n} = \hat{\mu}_{i,n} + \sqrt{\frac{1}{2n} \log \frac{2Kn^2}{\delta}},\ \underline{\mu}_{i,n} = \hat{\mu}_{i,n} - \sqrt{\frac{1}{2n} \log \frac{2Kn^2}{\delta}}, \tag{5}$$

where $\hat{\mu}_{i,n}$ is the sample mean of rewards of arm i after n pulls. We use the following notations as well for the sake of simplicity.

$$\overline{\mu}_i(t) = \overline{\mu}_{i,n_i(t+1)},\ \underline{\mu}_i(t) = \underline{\mu}_{i,n_i(t+1)}. \tag{6}$$

Here, $n_i(t+1)$ is the number of pulls of arm i after t pulls in total.

The decision condition for positiveness of arm i is $\underline{\mu}_i(t) \geq \theta$ and that for negativeness of arm i is $\overline{\mu}_i(t) < \theta$.

For these decision conditions, the following lemma guarantees the probability of wrong decision for each arm is at most δ/K.

Lemma 2. *For a positive arm i (i.e., $\mu_i \geq \theta$), $\overline{\mu}_i(t) \geq \theta$ holds for any time step t with probability at least $1 - \frac{\delta}{K}$. For a negative arm i (i.e., $\mu_i < \theta$), $\underline{\mu}_i(t) < \theta$ holds for any time step t with probability at least $1 - \frac{\delta}{K}$.*

Proof. For a positive arm i, we have

$$\begin{aligned}
\mathbb{P}[\exists t,\ \overline{\mu}_i(t) < \theta] =& \mathbb{P}[\exists n,\ \overline{\mu}_{i,n} < \theta] \leq \sum_{n=1}^{\infty} \mathbb{P}[\overline{\mu}_{i,n} < \theta] \\
=& \sum_{n=1}^{\infty} \mathbb{P}\left[\hat{\mu}_{i,n} + \sqrt{\frac{1}{2n}\log\frac{2Kn^2}{\delta}} < \theta\right] \\
\leq& \sum_{n=1}^{\infty} \mathbb{P}\left[\hat{\mu}_{i,n} < \mu_i - \sqrt{\frac{1}{2n}\log\frac{2Kn^2}{\delta}}\right] \ \text{(because } \mu_i \geq \theta\text{)} \\
\leq& \sum_{n=1}^{\infty} \exp\left(-2n\left(\sqrt{\frac{1}{2n}\log\frac{2Kn^2}{\delta}}\right)^2\right) \ \left(\begin{matrix}\text{by Hoeffding's} \\ \text{inequality}\end{matrix}\right) \\
=& \sum_{n=1}^{\infty} \frac{\delta}{2Kn^2} < \frac{\delta}{K} \ \left(\text{because } \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < 2\right).
\end{aligned}$$

Therefore, for a positive arm i, the probability that $\overline{\mu}_i(t) > \theta$ always holds for any time step t is larger than $1 - \frac{\delta}{K}$.

Similarly, we can show the inequality for a negative arm as well. □

As $n \to \infty$, $\hat{\mu}_{i,n}$ goes to μ_i from law of large numbers and $\sqrt{\frac{1}{2n}\log\frac{2Kn^2}{\delta}}$ goes to 0. Therefore, a positive arm i satisfies $\underline{\mu}_{i,n} > \theta$ (i.e. it is diagnosed as a positive arm) for some finite n if $\mu_i \neq \theta$. Here $\mu_i = \theta$ generally corresponds to the situation that infinite number of draw of the arm i is required for diagnosis. From Lemma 2, the probability that a positive arm i satisfies $\overline{\mu}_i(t_0) < \theta$ (i.e. it is diagnosed as a negative arm) is at most $\frac{\delta}{K}$. Therefore a positive arm is diagnosed as a positive arm correctly with failure probability at most $\frac{\delta}{K}$. Similarly, a negative arm is diagnosed correctly with failure probability at most $\frac{\delta}{K}$. Therefore the failure probability that the agent diagnoses any arm wrongly is at most δ as long as $\mu_i \neq \theta$ for any arm i. The agent counts the number of

positive arm n_P and negative arm n_N by using these conditions at each time step t, and stops and outputs "positive" when $n_P \geq \lambda$ or outputs "negative" when $K - n_N < \lambda$ since the number of positive arms cannot exceed $K - n_N$. Since the counts n_P and n_N are correct with probability at least $1 - \delta$ from the above discussion, Algorithm 1 solves one-threshold classification bandits with a specified confidence. We have proved the following theorem.

Theorem 2. *For one-threshold classification bandits with parameters $K, \theta, \lambda, \delta$, the outputs of Algorithm 1 using any arm selection policy satisfy the requirements of Problem 2.*

4.2 Arm Selection Policies

Let A_t be a set of arms that have not been diagnosed as positive or negative before time step t by the conditions explained in the previous section. It is enough to choose arm only from A_t at each time step t.

We developed the arm selection policy based on the Thompson sampling. Let θ_i be the parameter of the reward distribution ν_i of arm i. Assume a prior distribution π_i^0 of θ_i. The original Thompson sampling estimates a posterior distribution π_i^{t-1} of θ_i for each arm i at each time step t, and chooses an arm. The proposed algorithm is described as follows:

ThompsonSampling-CB:
1. For each arm $i \in A_t$,
(1) Calculate the posterior distribution π_i^{t-1} of θ_i using all the rewards obtained by time step t.
(2) Sample $\hat{\theta}_i \sim \pi_i^{t-1}$.
(3) Calculate the expected mean for given $\hat{\theta}_i$: $\tilde{\mu}_i^t = \mathbb{E}_{P[X|i,\hat{\theta}_i]}[X]$.
2. Count the number of arms i with $\tilde{\mu}_i^t$ at least θ, $B_t = |\{i \in [K] | \tilde{\mu}_i^t \geq \theta\}|$.
3. Select arm $i_t = \begin{cases} \arg\max_{i \in A_t} \tilde{\mu}_i^t & (\text{when } B_t \geq \lambda) \\ \arg\min_{i \in A_t} \tilde{\mu}_i^t & (\text{when } B_t < \lambda) \end{cases}$

For comparison, we examine the following arm selection policies as well.

UCB-CB:
Select $i_t = \arg\max_{i \in A_t} \hat{\mu}_i(t) + \sqrt{\frac{1}{2n_i(t)} \log t}$

Successive Elimination-CB:
Select $i_t = \arg\min_{i \in A_t} n_i(t)$

In both arm selection policies, if more than one argument satisfy the condition, one of them is chosen arbitrarily. As one of comparison methods, we select UCB algorithm because it is the best performer for good arm identification problem [4], whose problem setting is most similar to the setting of our one-threshold classification problem.

5 Experiments

In this section, we show the result of comparison experiments for the three algorithms that we proposed in Sect. 4.2.

The stopping time of Algorithm 1 using ThompsonSampling-CB is compared with those using UCB-CB and Successive Elimination-CB for the one-threshold classification bandits with positive δ that is reduced from classification bandits instances. In this experiment, we fixed parameters $K, \theta, p_{\mathrm{TN}}, p_{\mathrm{TP}}, \delta_P$ and δ_N of the original classification bandits instances as $K = 20$, $\theta = 0.5$, $p_{\mathrm{TN}} = p_{\mathrm{TP}} = 0.95$ and $\delta_P = \delta_N = 0.1$. Expected reward μ_i of arm i is taken from a uniform distribution over $[0, \theta]$ for negative arms and $[\theta, 1]$ for positive arms. The distributions of reward are Bernoulli distribution. For $N_1 = 0, 1, \ldots, 14$, $N_2 = N_1 + 5$ and the cases with just $5, 10, 15$ positive arms, we reduced each problem instance to the corresponding one-threshold problem with parameters λ and δ, and measured the stopping time (the number of samples) of Algorithm 1.

The results are shown in the upper left, the upper right and the lower left graphs for the cases with just $5, 10, 15$ positive arms, respectively, of Fig. 2. For

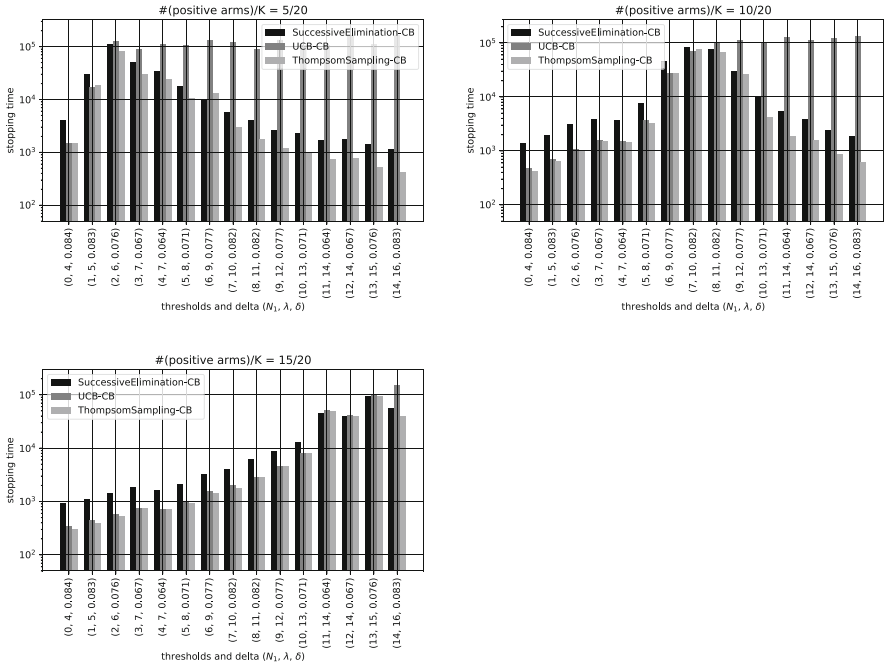

Fig. 2. Average stopping times (the number of samples) over 100 runs of Algorithm 1 using three arm selection policies for one-threshold classification bandits instances with parameters λ and δ reduced from classification bandits instances with parameters $K = 20$, $\theta = 0.5$, $p_{\mathrm{TN}} = p_{\mathrm{TP}} = 0.95$, $\delta_P = \delta_N = 0.1$, $N_1 = 0, 1, \ldots, 14$ and $N_2 = N_1 + 5$. y-axis is 'log' scale. The upper left, the upper right and the lower left graphs are results for the case with just $5, 10$, and 15 positive arms, respectively.

these three graphs, we can see ThompsonSampling-CB always stops earlier than SuccessiveElimination-CB and always stops earlier or in time comparable to UCB-CB. The performance of UCB-CB is poor when the output should be "negative" because UCB-CB tries to select positive arms with higher priority.

6 Conclusion and Future Works

In this paper, we presented an algorithm to reduce classification bandits problem based on an imperfect classifier with nonnegligible false-positive and false-negative rates into a one-threshold classification bandits problem under the allowed error rate δ. The parameters of true negative and positive probabilities p_{TN}, p_{TP}, and the number of arms K are supposed to be given in actual applications. Then the question here was whether we can still discriminate the number of bad arms is at most N_1 with probability at least $1 - \delta_N$ or at least N_2 $(> N_1)$ with probability at least $1 - \delta_P$. Usually confidence parameter δ required for bandits algorithms is the same as a given parameter itself, but in classification bandits, the confidence parameter δ required for the transformed one-threshold classification bandits is smaller than given parameters δ_N and δ_P for original classification bandits. Our reduction theorem enables us not only to provide the error rate δ smaller than originally given δ_P and δ_N but also to suggest whether it is difficult to find algorithm satisfying the given confidence level when $\delta < 0$.

For future work, we plan to apply our algorithm for classification bandits to interactive measurement by Raman microscope for differentiating cancer cells and non-cancer cells, where no one can identify whether each cell is cancer or not with 100% accuracy (at best 80–95% for example). Theoretically there exists some room to improve the algorithm such as selection policy although in our simulation through Thompson sampling was found to be superior in performance to the algorithms based on UCB and successive elimination.

References

1. Audibert, J.Y., Bubeck, S.: Best arm identification in multi-armed bandits (2010)
2. Gabillon, V., Ghavamzadeh, M., Lazaric, A.: Best arm identification: a unified approach to fixed budget and fixed confidence. In: Advances in Neural Information Processing Systems, pp. 3212–3220 (2012)
3. Kalyanakrishnan, S., Tewari, A., Auer, P., Stone, P.: Pac subset selection in stochastic multi-armed bandits. In: ICML, vol. 12, pp. 655–662 (2012)
4. Kano, H., Honda, J., Sakamaki, K., Matsuura, K., Nakamura, A., Sugiyama, M.: Good arm identification via bandit feedback. Mach. Learn. **108**(5), 721–745 (2019). https://doi.org/10.1007/s10994-019-05784-4
5. Kaufmann, E., Koolen, W.M., Garivier, A.: Sequential test for the lowest mean: from Thompson to murphy sampling. In: Advances in Neural Information Processing Systems, pp. 6332–6342 (2018)
6. Locatelli, A., Gutzeit, M., Carpentier, A.: An optimal algorithm for the thresholding bandit problem. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning, vol. 48, pp. 1690–1698 (2016)

7. Nachlas, J.A., Loney, S.R., Binney, B.A.: Diagnostic-strategy selection for series systems. IEEE Trans. Reliab. **39**(3), 273–280 (1990)
8. Pelissier, A., et al.: Intelligent measurement analysis on single cell Raman images for the diagnosis of follicular thyroid carcinoma. arXiv preprint (2019). arxiv.org/abs/1904.05675
9. Raghavan, V., Shakeri, M., Pattipati, K.: Test sequencing algorithms with unreliable tests. IEEE Trans. Syst. Man Cybern.-Part A: Syst. Humans **29**(4), 347–357 (1999)
10. Tabata, K., Nakamura, A., Honda, J., Komatsuzaki, T.: A bad arm existence checking problem: how to utilize asymmetric problem structure? Mach. Learn. **109**, 1–46 (2019). https://doi.org/10.1007/s10994-019-05854-7