With a Little Help from the Computer: Hybrid Human-Machine Systems on Bandit Problems

Bryan R. Gibson, Kwang-Sung Jun, Xiaojin Zhu
Department of Computer Science
University of Wisconsin-Madison, Madison, WI 53703
{bgibson, deltakam, jerryzhu}@cs.wisc.edu

Abstract

A common task for learners, both human and machine, is to choose from a set of actions with unknown reward distributions, with the objective of maximizing the reward over time. There are algorithms proven to perform optimally on this Multi-Arm Bandit task. A natural question is whether such algorithms can be used to enhance human performance in a human-machine hybrid system. We design and conduct a series of behavioral experiments to investigate this question.

1 Introduction

The process of choosing actions to maximize reward is a common task. Imagine a person sitting before two slot machines with unknown payoff distributions. The person has to decide which of them to put a coin into and play. This is known as a Multi-Arm Bandit (MAB) problem and has been well studied in both machine learning and psychology [1, 3, 5]. It is an example of the exploration-exploitation trade-off. If we reformulate the problem to one of minimizing regret, defined later, algorithms such as UCB1 described below have been designed to solve the task optimally [2]. On the other hand, humans are known to be suboptimal on MAB problems.

In many real world situations, it is the human that makes the final decision. Consider a setting where the machine algorithm (running on a wearable computer, for example) observes a human solving a MAB problem and gives suggestions to the human. However, the human can ignore the machine suggestions. Is it possible that such machine-human hybrid system performs better than humans would on their own? We propose a study which looks at that question.

1.1 Machine Learning Solution

We first review the UCB1 algorithm [2]. In all of the following we restrict ourselves to the special case of two arms with fixed but unknown distributions.

Let A and B be the unknown distributions belonging to two arms, with μ_A, μ_B their mean respectively. Let the reward be x. The UCB1 algorithm operates by using past experience to find the distribution which has the highest expected upper bound. At any iteration, it chooses arm

$$\arg\max_j \bar{x}_j + \sqrt{\frac{2\ln n}{n_j}} \tag{1}$$

to play, where n is the total number of iterations, n_j the number of iterations where arm j has been chosen, and $\bar{x}_j$ the average reward from arm j. We can now define 'optimal' as the action which chooses arm j corresponding to $\max\{\mu_A, \mu_B\}$.

While the UCB1 algorithm has been proven optimal with regret $O(\ln n)$, a modified version known as UCB1-Tuned (UCB1t) was found to work better empirically. The calculation used by UCB1t to choose the next draw is

$$\arg\max_j \bar{x}_j + \sqrt{\frac{\ln n}{n_j} \min\left(\frac{1}{4}, V_j(n_j)\right)} \tag{2}$$

where

$$V_j(s) = \left(\frac{1}{s}\sum_{\tau=1}^{s} x_{j\tau}^2\right) - \bar{x}_{js}^2 + \sqrt{\frac{2\ln t}{s}} \tag{3}$$

which gives a new upper bound using sample variance for arm j, which has been played s times in t iterations.

2 Human MAB Experiments

Participants were given two arms with differing distributions, each returning an integer reward $x_i \in [1, 100]$. The participant's task was to maximize their reward over a set number of iterations, or pulls. A UCB1t learner also detected which arm was pulled and the reward received, but could not pull arms itself. The only way the machine learner could affect the pulls was to communicate a suggestion to the human, who might or might not follow the suggestion.

Due to the fact that the human might not agree with the machine suggestion, the regret could still be higher than optimal. One novel aspect of our machine-human hybrid system is the following. In addition to learning the optimal pull, a machine learner could also learn to predict how likely the participant will agree with its suggestion. Let G_i be the event that the participant agrees with the machine suggestion on iteration i, x_i the reward at i, and S_i the machine suggestion at i. The learner tries to predict

$$P(G_i | G_{1:i-1}, x_{1:i-1}, S_{1:i-1}) \tag{4}$$

which is the probability of agreement given history. If the probability of agreement is lower than 1/2, the suggestion could be flipped in an attempt to manipulate the human into selecting the optimal decision by disagreeing; i.e., a 'reverse psychology' strategy.

2.1 Participants

112 university undergraduates participated for partial course credit.

2.2 Materials

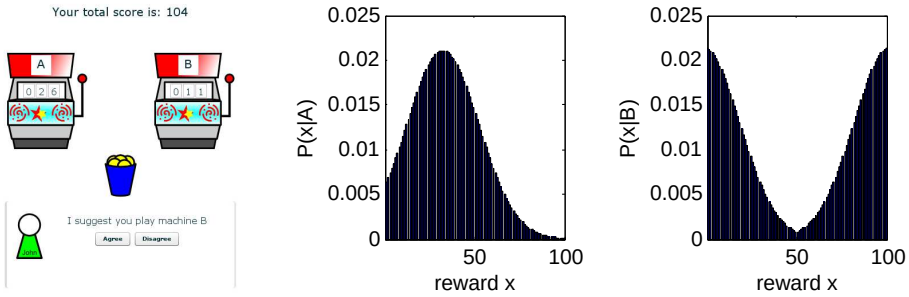
In our experiment, two distributions were designed to confuse the human learner into choosing a suboptimal strategy (Figures 1(b), 1(c)). Arm A used a bell shaped distribution with $\mu_A = 35.2$. Arm B used a distribution with the majority of its mass near the boundaries of its range but a mean value of $\mu_B = 50.5$. Draws from arm A would more consistently be just below the midpoint of the range while draws from B would vary widely between high and low rewards. The hope was that these low rewards would indicate to the participant that arm B was suboptimal even though it is in fact the optimal choice.

A computer interface was created to represent two arms, a display of total reward achieved and a suggestion display, as shown in Figure 1(a).

Participant interaction with the interface differed slightly by condition. In conditions where suggestions were given, a simple graphic of a person along with 'Agree' and 'Disagree' buttons were presented in the suggestion area. Participants chose which arm to play solely by clicking these buttons.

When no suggestions were given the suggestion area remained empty. To activate an arm the participant clicked on a representation of a 'coin bucket', followed by a click on one of the arms themselves. The 'coin bucket' click was implemented to keep participants from simply clicking the arm the mouse was closest to.

Once an arm was activated the reward amount was displayed on the arm itself with the display of total reward acquired updated as well.



(a) Interface (b) $\mu_A = 35.2$ (c) $\mu_B = 50.5$

Figure 1: The experimental interface and distributions.

2.3 Procedure

Each participant was given instructions indicating the range of reward, the fact that one arm might give better results than the other, that they might or might not receive suggestions about which arm to play, and the fact that their goal was to maximize their reward.

Participants then completed 150 iterations, each consisting of a single pull of an arm followed by the display of a reward and possibly a new suggestion depending on condition.

2.4 Conditions

Four conditions were chosen as simple examples of possible suggestion regimes:

H : No suggestions given. Participants interacted directly with the arms.

S : A simple suggestion was given of the form "I suggest you play machine A".

S+ : A more authoritative suggestion was given which included the statistics used by UCB1t to come to its decision. It took the form "You have played machine A (B respectively) 3 (5) times, the sample mean is 45 (72), while the upper confidence bound of the true mean can be as high as 87 (100). I suggest you play machine B."

RP : Before a simple suggestion of the same form used in **H** was displayed, the probability that the participant would agree with the suggestion given was calculated. A simple approximation to the model discussed above was used, conditioned solely on the agreement during the last iteration: $P(G_i|G_{i-1})$. Let M_i be the true intention of the machine learner and $\neg M_i$ the opposite. The suggestion S_i given was

$$S_i = \begin{cases} M_i & \text{if } P(G_i|G_{i-1}) \geq 1/2 \\ \neg M_i & \text{otherwise.} \end{cases} \tag{5}$$

In other words, if the probability of agreement on the current iteration was low, the machine learner attempted a reverse psychology strategy.

In all conditions where a suggestion was given, the first two suggestions were to pull arms that had not yet been played as the UCB1t learner is unable to make any predictions till at least one sample has been taken from both distributions.

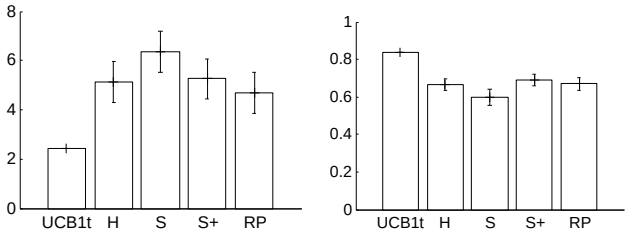
3 Results

Two metrics were used to measure performance, per-trial regret

$$\mu^* - \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

and a 'best-arm' percentage. Best-arm percentage is the percentage of total iterations where the optimal arm was pulled. As a comparison, UCB1t was run on its own for 5000 sessions, each

session consisting of 29 trials of 150 iterations. Figures 2(a) and 2(b) show the mean performance in each condition. The number of participants per conditions **H**, **S**, **S+** and **RP** was 28, 27, 28 and 29 respectively.



(a) Per-Trial Regret (b) Best Arm

Figure 2: Result means per condition.

The differences between human conditions are not statistically significant, however the trend is surprising in that suggestions do not seem to have helped and, in fact, have hurt in **S**. Performance is at best equivalent to human performance without suggestions.

4 Discussion

It is interesting to see that humans are not helped by such machine suggestions, given that the suggestions are optimal. Although it is a negative result, we believe our work is still valuable in that it provides a novel perspective on solving MAB problems with a machine-human hybrid system, where the machine plays the assistant but the human has the free will to choose. We speculate that our machine assistant can be more successful, if it can give suggestions in a form that is easier for humans to accept. For instance, it might suggest "You may want to try exploring different arms more" instead of a concrete arm suggestion at each iteration. A more complex model of participant agreement, taking into account the full history, might improve the performance of reverse psychology as well. Additionally, it may be informative to compare these results to a similar experiment using Gittins' Dynamic Allocation Process on a direct maximization of the reward [4].

Acknowledgments

Thank you to our reviewers for their helpful suggestions.

This research is supported in part by NSF IIS-0953219 and AFOSR FA9550-09-1-0313.

References

[1] D. Acuna and P. Schrater. Bayesian modeling of human sequential Decision-Making on the Multi-Armed bandit problem. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235256, 2002.

[3] Nathaniel D. Daw, John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006.

[4] J. C. Gittins and D. M. Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):pp. 561–565, 1979.

[5] M. Steyvers, M. D Lee, and E. J Wagenmakers. A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168179, 2009.