ESTUDO SOBRE MULTI-ARMED BANDITS

A STUDY OF MULTI-ARMED BANDITS

Diogo Cardoso Fernandes [1]
Mario Henrique Adaniya [2]

RESUMO

Este trabalho analisa as classificações do algoritmo multi-Armed bandits (MAB) e os seus campos de aplicação, por ser um algoritmo genérico o MAB tem diversas aplicações atuais. Um algoritmo MAB refere-se à uma subclasse do aprendizado por reforço onde o algoritmo, tem um agente que segue um processo sequencial de decisão em que procura otimizar suas ações enquanto melhora seu conhecimento sobre as opções disponíveis no ambiente. Sendo estas aplicações de problema bandidos encontradas em várias áreas de pesquisa envolvendo desde sistemas de recomendações, teste clínicos, problemas de roteamento em redes cognitivas, pesquisa operacional, economia e estatística. O objetivo deste trabalho é realizar um estudo dos problemas bandidos e suas diferentes configurações, e encontrar na literatura quais passos serão abordados em trabalhos futuros.

Palavras-chave: multi-armed bandit; contextual bandit; reinforcement learning.

ABSTRACT

This work analyzes the classifications of the multi-Armed bandits (MAB) algorithm and its fields of application, as it is a generic algorithm, MAB has several current applications. A MAB algorithm refers to a subclass of reinforcement learning where the algorithm has an agent that follows a sequential decision process. in which it seeks to optimize its actions while improving its knowledge of the options available in the environment. These bandit problem applications are found in several research areas ranging from recommender systems, clinical tests, routing problems in cognitive networks, operational research, economics, and statistics. The objective of this work is to conduct a study of bandit problems and their different configurations, and to find in the literature which steps will be addressed in future works.

Keywords: multi-armed bandit; contextual bandit; reinforcement learning.

[1] Graduando do Curso de Ciência da Computação do Centro Universitário Filadélfia - UniFil. fdiogoc@edu.unifil.br
[2] Orientador: Professor Mario Henrique Adaniya do Curso de Ciência da Computação do Centro Universitário Filadélfia - UniFil. mario.adaniya@unifil.br

1 INTRODUÇÃO

De acordo com Goldschmidt (2010) a Inteligência Computacional vai além da perspectiva de compreensão do pensamento humano, pois procura construir entida des artificiais inteligentes. Segundo o autor algumas habilidades que necessariamente envolvem o conceito de inteligência são:

- Capacidade de raciocínio / dedução / inferência;
- Capacidade de aprendizado;
- Capacidade de percepção;
- Capacidade de evolução e adaptação.

Conforme relata Sutton e Barto (2018), a aprendizagem por reforço é uma abordagem computacional que propõe entender e automatizar a aprendizagem e tomada de decisão direcionadas a objetivos. Distingue-se das outras abordagens por sua ênfase na aprendizagem por um agente interagindo diretamente com o ambiente, sem exigir supervisão ou modelos do ambiente. o aprendizado por reforço é o primeiro campo a abordar seriamente os problemas computacionais que surgem ao aprender a partir da interação com um ambiente com foco em atingir objetivos de longo prazo.

O agente precisa interagir com o ambiente e observar as respostas do ambiente pois precisa estimar uma política reforçando suas crenças sobre a dinâmica do ambiente. Com o tempo, o agente passa a entender como o ambiente responde às suas ações, podendo assim começar a estimar uma política ótima. No aprendizado por reforço, o agente estima uma política ótima para se comportar em um ambiente desconhecido (ou parcialmente conhecido) interagindo com ele (usando uma abordagem de "tentativa e erro").

Sutton e Barto (2018) aponta que no aprendizado por reforço, a exploração não precisa se limitar a realizar ações desconhecidas, as ações podem ser geradas por métodos sofisticados usando conhecimento previamente aprendido, desde que haja alguma exploração.

Os processos de aprendizagem estão sujeitos a algumas limitações importantes. A aprendizagem tem que lidar com a experiência confusa e problemática de equilibrar os objetivos concorrentes. Objetivos esses de desenvolver novos conhecimentos (exploring) e explorar (exploiting) a sua situação atual em face de

tendências dinâmicas do ambiente para escolher entre exploring-exploiting.

Multi-armed bandits (MAB) é um framework genérico para problemas de tomada de decisão sequenciais que busca a otimização de recompensa desconhecida. Onde precisamos encontrar um meio termo entre explorar o melhor caminho ou descobrir um novo caminho com uma possível melhor recompensa. O objetivo do algoritmo bandido é encontrar uma sequência de ações que maximize a soma de recompensas durante o experimento, minimizando o arrependimento total, sendo arrependimento a diferença entre a recompensa obtida e a que seria obtida utilizando uma política ótima.

No mundo atual tomar decisões do tipo exploring-exploiting é uma situação recorrente, e nós humanos fazemos isto várias vezes por dia e certa com facilidade (WILSON *et al*., 2014).

Problemas MAB são estudados desde (THOMPSON, 1933), podemos encontrar o dilema entre explorar-extrair, em várias áreas de pesquisa envolvendo desde sistemas de recomendações, teste clínicos, problemas de roteamento em redescog nitivas, pesquisa operacional, economia, estatística (ZHOU, 2015) (SLIVKINS, 2021).

A exploração ativa de informação considerando a ausência de um contexto anterior é que torna os problemas MAB um problema de aprendizado online, portanto um subset do aprendizado por reforço.

2 DESENVOLVIMENTO

Existem duas abordagens sobre como tirar proveito da experiência gerada pelo aprendizado por reforço: Model-Based (MB) e Model-Free (MF) a distinção entre algoritmos MF e MB corresponde à distinção que os psicólogos fazem entre controle habitual e controle direcionado a objetivos. Hábitos são padrões de comportamento despertados por estímulos apropriados e então executados mais ou menos automaticamente.

O comportamento direcionado a objetivos, de acordo com a forma como os psicólogos usam a frase, é proposital no sentido de que é controlado pelo conhecimento do valor dos objetivos e da relação entre as ações e suas

consequências. Hábitos são, às vezes, controlados por estímulos antecedentes, enquanto comportamentos direcionados a objetivos são controlados por suas consequências.

O controle direcionado a objetivos tem a vantagem de poder mudar rapidamente o comportamento quando o ambiente muda sua maneira de reagir às ações. Embora o comportamento habitual responda rapidamente à entrada de um ambiente acostumado, ele é incapaz de se ajustar rapidamente às mudanças no ambiente (GLÄSCHER *et al*., 2010; BANSAL *et al*., 2017).

Segundo Deisenroth, Neumann e Peters (2013) as abordagens de pesquisas MB e MF têm sido desenvolvidas em sua maioria isoladamente. No entanto, a combinação de políticas com modelos de busca MF e com modelos MB parece ser uma abordagem promissora.

Uma política MB que explora gananciosamente um modelo aprendido pode usar uma política MF para atualizar sua estratégia evitando uma escolha sub ótima local, e, portanto, melhorando a qualidade do modelo aprendido.

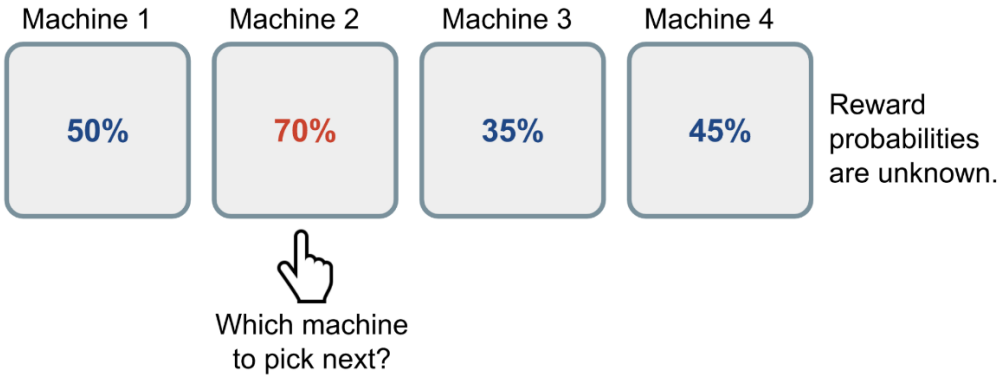
Ainda em Deisenroth, Neumann e Peters (2013) vemos que uma outra abordagem é trocar entre políticas MB e MF durante o processo de aprendizado, utilizando uma política MB para explorar o ambiente e coletar informações iniciais suficiente para que uma política MF seja executada com eficiência, evitando assim que erros no modelo inicial afetem o resultado do algoritmo.

Multi-armed bandits (MAB) refere-se à uma subclasse do aprendizado por reforço onde o algoritmo, tem um agente que segue um processo sequencial de decisão em que procura otimizar suas ações enquanto melhora seu conhecimento sobre as opções disponíveis no ambiente.

Imagine um jogador, apostando em mais de uma máquina caça-níquel com uma probabilidade de recompensa desconhecida. O principal objetivo do jogador é maximizar os seus ganhos. Se o jogador soubesse a recompensa de cada máquina ele somente jogaria na que tem a maior probabilidade de recompensa.

Na Figura 1 temos 4 máquinas cada uma com um retorno fixo, porém desconhecido pelo jogador.

Figura 1 – Ilustração de um multi-armed bandit Bernoulli.



Fonte: Weng (2018)

Como o jogador não sabe qual a recompensa é preciso jogar em todas para saber qual a que trará o maior retorno. A recompensa das máquinas pode não ser fixa então o jogador precisa jogar mais de uma vez em cada para que ele tenha uma melhor visão da recompensa média de cada máquina caça níquel. Tendo de encontrar balanceamento entre a que retornou a maior recompensa no passado (exploitation) ou escolher outra que pode retornar uma recompensa maior (exploration). Foi deste cenário em uma máquina caça-níquel que surgiu o termo MAB.

Os “braços” neste problema podem representar diferentes produtos que podem ser exibidos em um site. Os usuários que chegam ao site são mostrados versões do site com diferentes produtos. Um sucesso está associado seja com um clique no anúncio ou com uma conversão (uma venda do item sendo anunciado). Os parâmetros θk representam a taxa de cliques ou a taxa de conversão entre os usuários que frequentam o site. o algoritmo espera equilibrar entre exploration e exploitation para maximizar o número total de sucessos possível.

Na Tabela 1 abaixo temos exemplificados mais diversos domínios que podemos usar um algoritmo bandido para ajudar a encontrar um balando entre exploration-exploitation.

Tabela 1 – Domínios de aplicações MAB.

Domínio da aplicação	Ação	Recompensa
Estudos clínicos	qual droga prescrever	Melhora da saude
Web design	e.g. cor da fonte ou layout da página	$\sum$ cliques
Otimização de cont.	quais itens enfatizar	$\sum$ cliques
Pesquisa na web	itens a priorizar a pequisa	$\sum$ usuários satisfeitos
Anúncio	qual anúncio mostrar	retorno financeiro
Sis. de recomendação	e.g. qual filme assistir	1 se seguir recomendação
Otimização de vendas	quais produtos e valores mostrar	retorno financeiro
Leilão	e.g. qual preço de reserva utilizar	retorno financeiro
Crowdsourcing	match entre tarefas e trabalhadores	$\sum$ tarefas completas
Internet	qual configuração TCP utilizar	qualidade da conexão
Redes cognitivas	qual frequência de rádio utilizar	$\sum$ transmissões
Controles robóticos	e.g. qual estratégia utilizar	tempo para completar tarefa
Banco de dados	e.g. qual index criar	tempo para completar tarefa
Jogos como adversário	e.g. qual estratégia utilizar	vitória
Jogos como parceiro	e.g, qual estratégia utilizar	satisfação do jogador
Hiper heurísticas	qual heurística de baixo nível utilizar	qualidade da resposta

Fonte: (SLIVKINS, 2021) (Editado e traduzido).

2.1 ESTRATÉGIAS PARA SELEÇÃO DE AÇÃO POR UM MAB

Conforme o algoritmo define como irá explorar (exploration) o ambiente, existem diferentes modos de se resolver o problema MAB. Pode-se optar por uma estratégia ruim como nunca utilizar exploration, pode-se utilizar exploration de maneira randômica, ou pode-se criar uma estratégia de exploration que de preferência a dados desconhecidos.

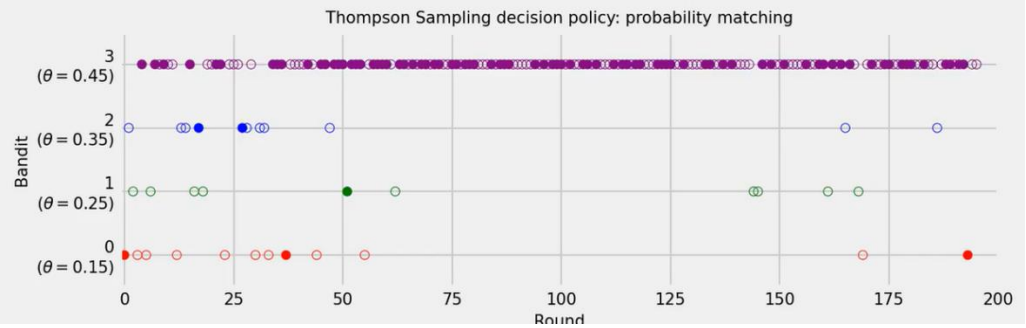
Weng (2018) cita que uma abordagem ingênua para este problema envolve alocar períodos para exploração e em cada período escolher um braço uniformemente ao acaso, visando selecionar ações bem-sucedidas em outros períodos de tempo. Tal abordagem pode ser bastante desperdício, para um simples problema do bandido Bernoulli pode falhar completamente para problemas mais complicados.

2.1.1 Thompson Sampling

Thompson samplin (TS) é um algoritmo que traz a inferência bayesiana para o MAB (CHAPELLE; LI, 2011), a ideia do algoritmo TS é de randomicamente escolher cada braço de acordo com a sua probabilidade de ser ótimo. Em sua modulação mais simples o algoritmo TS não tem parâmetros para serem modificados tornando-se

simples de ser implementado, como é um algoritmo aleatório acaba se tornando seguro para utilização em ambientes com retorno do estado e recompensa atrasados.

Figura 2 – Experimento com política TS contendo quatro braços e duzentos turnos.



Fonte: Marmerola (2017)

Na Figura 10 temos quatro braços com probabilidade θk ∈ [0, 1] a política TS, podemos ver que TS realiza uma exploração eficiente, descartando rapidamente braços menos promissores, mas não com muita rapidez. Braços menos promissores com alta incerteza são escolhidos, pois não existe informação suficiente para descartá-los. No entanto, quando a distribuição do melhor braço se destaca, considerando a incerteza, ficamos muito mais agressivos na exploração de recompensas desconhecidas (exploitation).

2.2 ALGORITMOS MAB CONTEXTUAIS PARA RECOMENDAÇÃO DE ANÚNCIOS

Os sistemas de recomendação de anúncios são alimentados por algoritmos de classificação e relevância e fornecem recomendações personalizadas. Isso ajuda os clientes a descobrir os restaurantes/marcas relevantes e vice-versa ajuda os parceiros de anúncios a atingir o público-alvo apropriado. Esses algoritmos de classificação geralmente são modelos de aprendizado de máquina treinados em dados históricos de interação e transação do cliente e otimizados para as taxas de cliques.

Sistemas de recomendação geralmente tendem a se sair bem ao recomendar parceiros de anúncios com os quais o cliente interagiu anteriormente ou aqueles que são semelhantes a esses parceiros. Mas eles normalmente não tendem a promover a descoberta de novos parceiros, que é um dos princípios básicos de qualquer plataforma de publicidade. Quando esses dados de interação são usados para

retreinar os modelos de ML, eles tendem a reforçar recomendações semelhantes devido ao viés de apresentação.

Uma das maneiras de quebrar esse círculo vicioso de preconceito e entender as preferências inexploradas dos clientes é por meio da coleta de dados imparcial por meio de recomendações aleatórias. Mas isso pode levar a uma experiência ruim do cliente entre o grupo de clientes expostos a essas recomendações aleatórias.

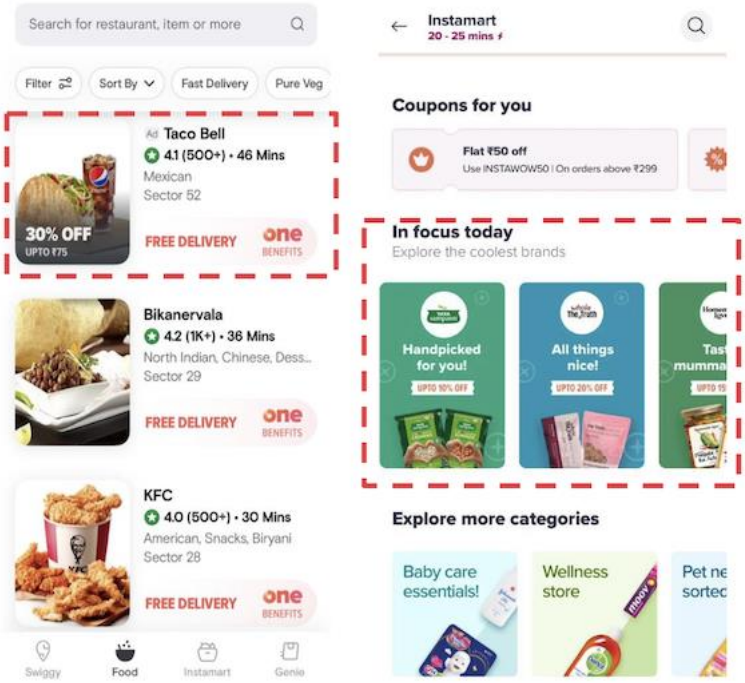
Um sistema de recomendação baseado em MAB recomendaria um parceiro de anúncio específico para um determinado cliente. Uma vez que a escolha é feita, o algoritmo observa a recompensa.

Quando o cliente fizer login no aplicativo da próxima vez, o algoritmo deve escolher o mesmo parceiro de anúncio porque o cliente clicou anteriormente ou o algoritmo deve tentar um novo parceiro de anúncio para obter novas informações?

Os algoritmos MAB fazem essa troca de tal forma que tentam aprender ou estimar a melhor escolha enquanto gasta o mínimo de tentativas explorando as opções.

Na versão básica de um algoritmo MAB, os braços escolhidos a cada vez são baseados puramente nas recompensas históricas que foram recebidas. Mas em muitos sistemas, pode existir informações ou contextos adicionais disponíveis antes de fazer a escolha. Por exemplo, informações sobre alguns perfis de gosto do cliente, preferências de preço etc. Os algoritmos de MABs contextuais usam essas informa ções adicionais e recompensas históricas para escolher o melhor braço. NA figura 3, destacamos onde MABs contextuais podem ser aplicados para recomendação de anúncios de restaurantes e ordenação de sugestões.

Figura 3 – Página inicial de um aplicativo.



Fonte: Bhavi; Shedthikere; Sunder (2022)

3 CONCLUSÃO

Neste trabalho, buscamos apresentar e discutir os conceitos fundamentais que embasam a área de Aprendizado por Reforço, tema central discutido, pois algumas abordagens de MAB estudadas investigam como otimizar o resultado e a sua aplicação em diversas áreas. Muitos desses algoritmos implementados possuem diversas limitações, sendo cada algoritmo focado em um problema específico.

Para trabalhos futuros foi escolhido utilizar algoritmos MAB baseados em TS, já que pela fácil modelagem do seu modelo de recompensa com distribuições normais, podemos aplicar este tipo de abordagem para diversos tipos de problema presentes em uma aplicação de celular focada em e-commerce, e oferecer personalizações ao usuário final a partir de decisões tomadas diretamente no próprio celular, na borda, sem que as informações precisem passar por um processamento externo e de alto custo.

REFERÊNCIAS

BANSAL, S. *et al*. *MBMF: Model-Based Priors for Model-Free Reinforcement Learning*. arXiv, 2017. Disponível em: https://arxiv.org/abs/1709.03153.

BHAVI, C.; SHEDTHIKERE, S.; SUNDER, S. The multi-armed bandit problem and its solutions. *https://bytes.swiggy.com*, 2022. Disponível em: https://bytes.swiggy.com/ contextual-bandits-for-ads-recommendations-ec210775fcf.

CHAPELLE, O.; LI, L. An empirical evaluation of thompson sampling. In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2011. (NIPS'11), p. 2249–2257.

DEISENROTH, M. P.; NEUMANN, G.; PETERS, J. [S.l.: s.n.], 2013.

GLÄSCHER, J. *et al*. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, Cell Press, v. 66, n. 4, p. 585–595, maio 2010.

GOLDSCHMIDT, R. R. *Inteligência Computacional*. [S.l.]: IST-Rio, 2010.

MARMEROLA, G. D. Introduction to thompson sampling: the bernoulli bandit. https://gdmarmerola.github.io/, 2017. Disponível em: https://gdmarmerola.github.io/ts-for-bernoulli-bandit/.

SLIVKINS, A. *Introduction to Multi-Armed Bandits*. 2021. Disponível em: https://arxiv.org/abs/1904.07272.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. Second. The MIT Press, 2018. Disponível em: http://incompleteideas.net/book/the-book-2nd.html.

THOMPSON, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 25, n. 3/4, p. 285–294, 1933. Disponível em: http://www.jstor.org/stable/2332286.

WENG, L. The multi-armed bandit problem and its solutions. *lilianweng.github.io*, 2018. Disponível em: https://lilianweng.github.io/posts/2018-01-23-multi-armed-bandit/.

WILSON, R. C. *et al*. Humans use directed and random exploration to solve the explore-exploit dilemma. *J. Exp. Psychol. Gen.*, v. 143, n. 6, p. 2074–2081, dez. 2014.

ZHOU, L. A survey on contextual multi-armed bandits. ago. 2015.