

Etat de l’art sur l’application des bandits multi-bras

1st Djallel Bouneffouf

IBM Research

Djallel.bouneffouf@ibm.com

New York, USA

Abstract—Le domaine des bandits multi-bras connaît actuellement une renaissance, alors que de nouveaux paramètres de problèmes et des algorithmes motivés par diverses applications pratiques sont introduits, en s’ajoutant au problème classique des bandits. Cet article vise à fournir un examen complet des principaux développements récents dans de multiples applications réelles des bandits. Plus précisément, nous introduisons une taxonomie des applications communes basées sur le MAB et résumons l’état de l’art pour chacun de ces domaines. De plus, nous identifions les tendances actuelles importantes et fournissons de nouvelles perspectives concernant l’avenir de ce domaine en plein essor.

Index Terms—bandit

I. INTRODUCTION

De nombreuses applications pratiques nécessitent des problèmes de prise de décision séquentiels, où un agent doit choisir la meilleure action parmi plusieurs alternatives. Des exemples de telles applications incluent les essais cliniques [1], systèmes de recommandation [2]–[11], [11]–[19], [19]–[45], [45]–[92] et la détection d’anomalies [93]. Dans certains cas, des informations secondaires ou un contexte sont associés à chaque action (par exemple, le profil d’un utilisateur), et le retour d’information, ou récompense, est limité à l’option choisie. Par exemple, dans les essais cliniques, le contexte est le dossier médical du patient (par exemple, état de santé, antécédents familiaux, etc.), les actions correspondent aux options de traitement comparées et la récompense représente le résultat du traitement proposé (par exemple, succès ou échec). Un aspect important affectant le succès à long terme dans de tels contextes est de trouver un bon compromis entre l’exploration (par exemple, essayer un nouveau traitement) et l’exploitation (choisir le traitement le plus connue à ce jour).

Ce compromis inhérent entre l’exploration et l’exploitation existe dans de nombreux problèmes de prise de décision séquentiels, et est traditionnellement

formulé comme le problème des bandit, qui se présente comme suit: Étant donné K actions possibles, ou “bras”, chacun associé à une distribution de probabilité de récompense fixe mais inconnue [94], [95], à chaque itération, un agent sélectionne un bras à jouer et reçoit une récompense, échantillonnée à partir de la distribution de probabilité du bras respectif indépendamment des actions précédentes. La tâche d’un agent est d’apprendre à choisir ses actions afin que les récompenses cumulées au fil du temps soient maximisées.

Notez que l’agent doit essayer différentes bras pour apprendre leurs récompenses (c’est-à-dire explorer le gain), et également utiliser ces informations apprises afin de recevoir le meilleur gain (exploiter les gains appris). Il existe un compromis naturel entre l’exploration et l’exploitation. Par exemple, essayer chaque bras exactement une fois, puis jouer le meilleur d’entre. Cette approche est souvent susceptible de conduire à des solutions très sous-optimales lorsque les récompenses des bras sont incertaines. Différentes solutions ont été proposées pour ce problème, basées sur une formulation stochastique [94]–[96] et une formulation bayésienne [97]; cependant, ces approches ne tenaient pas compte du contexte ou des informations secondaires dont disposait l’agent.

Il est à noter que le problème des bandits peut être vu comme la forme la plus simple d’apprentissage par renforcement, dans laquelle l’agent est sans état. Lorsque le système a des états, les actions provoquent des changements d’états et les récompenses dépendent également des états. Par conséquent, dans l’apprentissage par renforcement, les récompenses à différentes étapes ne sont pas indépendantes les unes des autres. En fait, les algorithmes classiques pour l’apprentissage par renforcement (avec états) utilisent souvent des solutions au problème des bandits multi-bras comme sous-programmes pour définir des politiques exploration exploitation dans l’apprentissage par renforcement. Par ex-

Identify applicable funding agency here. If none, delete this.

emple, il est bien connu ϵ -greedy L'algorithme de bandits multi-bras est souvent combiné avec l'algorithme de programmation dynamique de Bellman pour l'apprentissage par renforcement afin de définir les choix d'actions. En outre, de nombreux algorithmes d'apprentissage par renforcement, lorsqu'ils sont appliqués à des systèmes sans état, se réduisent à des algorithmes de bandit multi-bras.

Une version particulièrement utile du MAB est le contextual multi-arm bandit (CMAB), ou simplement le contextual bandit, où à chaque itération, avant de choisir un bras, l'agent observe un N -dimensions du contexte, ou vecteur de features. L'agent utilise ce contexte, ainsi que les récompenses des bras jouées dans le passé, pour choisir quel bras jouer dans l'itération actuelle. Au fil du temps, le but de l'agent est de collecter suffisamment d'informations sur la relation entre les vecteurs de contexte et les récompenses, afin qu'il puisse prédire le prochain meilleur bras à jouer en regardant le contexte actuel [98], [99]. Différents algorithmes ont été proposés pour le cas général, dont LINUCB [100], Neural Bandit [101] et Contextual Thompson Sampling (CTS) [99], où une dépendance linéaire est généralement supposée entre la récompense attendue d'une action et son contexte.

Nous allons maintenant fournir un aperçu des diverses applications du bandit au problèmes de la vie réelle (santé, réseau informatique, finance, et au-delà), ainsi qu'en apprentissage automatique. En particulier, lorsque les approches bandit peuvent aider à améliorer le réglage des hyperparamètres et d'autres choix algorithmiques importants dans l'apprentissage supervisé, l'apprentissage actif et l'apprentissage par renforcement.

II. APPLICATIONS DES BANDITS

Le bandit stochastique aborde les défis associés à la présence d'incertitude dans la prise de décision séquentielle. Ce type d'incertitude a une interaction complexe avec le dilemme de l'exploration exploitation et fournit donc un formalisme naturel pour la plupart des problèmes de prise de décision.

A. Santé

Essais cliniques. La collecte de données pour évaluer l'efficacité du traitement sur des animaux pendant tous les stades de la maladie peut être difficile lors de l'utilisation de procédures conventionnelles d'allocation de traitement aléatoire, car de mauvais choix de traitements peuvent entraîner une détérioration de la santé du sujet. Les auteurs de [1] visent à concevoir une stratégie d'allocation adaptative pour

améliorer l'efficacité de la collecte de données en allouant plus d'échantillons pour explorer des traitements prometteurs. Ils présentent cette application comme un problème de bandit contextuel et introduisent un algorithme pratique d'exploration exploitation dans ce cadre. Le travail repose sur le sous-échantillonnage pour comparer les options de traitement en utilisant une quantité équivalente d'informations. Ils étendent la stratégie de sous-échantillonnage au contexte de bandit contextuel en appliquant un sous-échantillonnage dans la régression avec processus gaussien.

Warfarine est l'anticoagulant oral le plus utilisé dans le monde; cependant, l'administration d'un dosage précis reste un défi important, car le dosage approprié peut être très variable entre les individus en raison de divers facteurs cliniques, démographiques et génétiques. Les médecins suivent actuellement une stratégie à dose fixe: les patients commencent avec une dose de 5 mg / jour (ce qui est la posologie appropriée pour la majorité des patients) et ajustent lentement la dose au cours de quelques semaines en suivant les taux d'anticoagulant du patient. Cependant, une posologie initiale incorrecte peut entraîner des conséquences très néfastes telles qu'un accident vasculaire cérébral (si la dose initiale est trop faible) ou une hémorragie interne (si la dose initiale est trop élevée). Ainsi, les auteurs de [102] abordent le problème de l'apprentissage et de l'attribution d'un dosage initial approprié aux patients en modélisant le problème comme un bandit avec des covariables de haute dimension, et proposent un nouvel algorithme de bandit efficace basé sur l'estimateur LASSO.

Modélisation du cerveau et du comportement. S'inspirant des études comportementales de la prise de décision humaine chez les patients souffrant de différents troubles mentaux, les auteurs de [43] proposent un cadre paramétrique général pour le problème des bandits qui étend l'approche standard d'échantillonnage de Thompson pour incorporer les biais de traitement des récompenses associés à plusieurs conditions neurologiques et psychiatriques, y compris les maladies de Parkinson et d'Alzheimer, le trouble de déficit de l'attention/ hyperactivité (TDAH), la dépendance et la douleur chronique. Ils démontrent empiriquement, du point de vue de la modélisation comportementale, que leur modèle peut être considéré comme une première étape vers un modèle de calcul unificateur capturant les anomalies du traitement des récompenses dans plusieurs conditions mentales.

B. La finance

Ces dernières années, la sélection séquentielle de portefeuilles a suscité un intérêt croissant à l'intersection de l'apprentissage automatique et de la finance quantitative. Le compromis entre l'exploration et l'exploitation, dans le but de maximiser la récompense cumulative, est une formulation naturelle des problèmes de choix de portefeuille. Dans [103], les auteurs ont proposé un algorithme de bandit pour faire des choix de portefeuille en ligne en exploitant les corrélations entre plusieurs bras. En construisant des portefeuilles orthogonaux à partir de plusieurs actifs et en intégrant leur approche au cadre des bandits, les auteurs dérivent la stratégie de portefeuille optimale représentant une combinaison d'investissements passifs et actifs selon une fonction de récompense ajustée au risque. Dans [104], les auteurs intègrent la conscience du risque dans le cadre classique du bandit et introduisent un nouvel algorithme pour la construction de portefeuille. En filtrant les actifs en fonction de la structure topologique du marché financier et en combinant la politique optimale de bandit avec la minimisation d'une mesure de risque, ils parviennent à un équilibre entre le risque et le rendement.

C. Tarification dynamique

Les entreprises de vente en ligne sont souvent confrontées au problème de tarification dynamique: l'entreprise doit décider des prix en temps réel pour chacun de ses multiples produits. L'entreprise peut mener des expériences de prix (faire des changements de prix fréquents) pour se renseigner sur la demande et maximiser les profits à long terme. Les auteurs de [105] proposent une politique d'expérimentation dynamique des prix, où l'entreprise ne dispose que d'informations incomplètes sur la demande. Pour ce paramètre général, les auteurs dérivent un algorithme de tarification qui équilibre le fait de gagner un profit immédiat par rapport à l'apprentissage pour les bénéfices futurs. L'approche combine un bandit avec une identification partielle de la demande des consommateurs à partir de la théorie économique. Semblable à [105], les auteurs de [106] considèrent la tarification multi-produits dynamique de haute dimension avec un modèle de demande linéaire de faible dimension évolutif. Ils montrent que le problème de maximisation des revenus se réduit à une optimisation convexe de bandit en ligne avec des informations secondaires données par les demandes observées. L'approche applique un algorithme d'optimisation convexe de bandit dans un espace projeté de faible dimension couvert par les caractéristiques du produit latent, tout en apprenant simultanément cette durée via

la décomposition en valeur singulière en ligne d'une matrice contenant les demandes observées.

D. Systèmes de recommandation

Les systèmes de recommandation sont fréquemment utilisés dans diverses applications pour prédire les préférences de l'utilisateur. Cependant, ils sont également confrontés au dilemme exploration-exploitation lorsqu'ils font une recommandation, car ils doivent exploiter leurs connaissances sur les éléments précédemment choisis qui intéressent l'utilisateur, tout en explorant de nouveaux éléments susceptibles de plaire à l'utilisateur. Les auteurs de [107] abordent ce défi en utilisant le paramètre bandit, en particulier pour les systèmes de recommandation à grande échelle qui ont un nombre vraiment grand ou infini d'éléments. Ils proposent deux approches de bandit à grande échelle dans des situations où aucune information préalable n'est disponible. Une exploration continue de leurs approches peut résoudre le problème du démarrage à froid dans les systèmes de recommandation. Dans les systèmes de recommandation contextuels, la plupart des approches existantes se concentrent sur la recommandation d'éléments pertinents aux utilisateurs, en tenant compte des informations contextuelles, telles que l'heure, le lieu ou les aspects sociaux. Cependant, aucune de ces approches n'a pris en compte le problème de l'évolution du contenu des utilisateurs. Dans [87], les auteurs introduisent un algorithme qui prend en compte cette dynamique. Il est basé sur une exploration / exploitation dynamique et peut équilibrer de manière adaptative les deux aspects, en décidant quelle situation est la plus pertinente pour l'exploration ou l'exploitation. En ce sens, [27] propose d'étudier la "fraîcheur" du contenu de l'utilisateur à travers le problème du bandit. Ils introduisent l'algorithme Freshness-Aware Thompson Sampling pour la recommandation de nouveaux documents.

E. Maximisation de l'influence

Les auteurs de [108] considèrent la maximisation de l'influence (IM) dans les réseaux sociaux, qui est le problème de maximiser le nombre d'utilisateurs qui prennent conscience d'un produit en sélectionnant un ensemble d'utilisateurs auxquels exposer le produit. Ils proposent une nouvelle paramétrisation qui rend non seulement le cadre indépendant du modèle de diffusion sous-jacent, mais aussi statistiquement efficace pour apprendre des données.

Ils donnent une fonction de substitution monotone et submoduleaire correspondante, et montrent qu'il s'agit

d'une bonne approximation de l'objectif original de la MI. Ils considèrent également le cas d'un nouveau marketeur cherchant à exploiter un réseau social existant, tout en apprenant simultanément les facteurs régissant la propagation de l'information. Pour cela, ils développent un algorithme de bandit basé sur LinUCB. Les auteurs de [109] étudient également le problème de maximisation de l'influence en ligne dans les réseaux sociaux mais sous le modèle de cascade indépendant. Plus précisément, ils essaient d'apprendre l'ensemble des "meilleures graines ou influenceurs" dans un réseau social en ligne tout en interagissant à plusieurs reprises avec lui. Ils abordent les défis de l'espace d'action combinatoire, car le nombre d'ensembles d'influenceurs réalisables augmente de manière exponentielle avec le nombre maximum d'influenceurs et un retour limité, car seule la partie influencée du réseau est observée.

F. Récupération de l'information

Les auteurs de [110] soutiennent que le processus de sélection itérative de recherche d'informations peut être naturellement modélisé comme un problème de bandit contextuel. Le modèle de bandit conduit à des méthodes très efficaces pour l'arbitrage des documents. Dans ce cadre d'attribution des bandits, ils proposent sept nouvelles méthodes de jugement de documents, dont cinq sont des méthodes stationnaires et deux sont des méthodes non stationnaires. Cette étude comparative comprend les méthodes existantes conçues pour l'évaluation basée sur la mise en commun et les méthodes existantes conçues pour la méta-recherche. Dans la recherche d'informations mobiles, les auteurs de [16] introduisent un algorithme qui aborde ce dilemme dans le domaine de la recherche d'informations basées sur le contexte (CBIR). Il est basé sur une exploration / exploitation dynamique et il peut équilibrer de manière adaptative les deux aspects en décidant quelle situation d'utilisateur est la plus pertinente pour l'exploration ou l'exploitation. Dans un cadre en ligne délibérément conçu, ils effectuent des évaluations auprès des utilisateurs mobiles.

G. Systèmes de Dialogue

Sélection de réponse de dialogue. La sélection de réponse de dialogue est une étape importante vers la génération de réponse naturelle dans les agents conversationnels. Les travaux existants sur les modèles conversationnels se concentrent principalement sur l'apprentissage supervisé hors ligne à l'aide d'un large ensemble de paires contexte-réponse. Dans [111], les auteurs se concentrent sur l'apprentissage en ligne de

la sélection des réponses dans les systèmes de dialogue. Ils proposent un modèle de bandit contextuel avec une fonction de récompense non linéaire qui utilise une représentation distribuée du texte pour la sélection de réponse en ligne. Un LSTM bidirectionnel est utilisé pour produire les représentations distribuées du contexte de dialogue et des réponses, qui servent d'entrée à un bandit contextuel. Ils proposent une méthode d'échantillonnage personnalisée de Thompson qui est appliquée à un espace de caractéristiques polynomiales pour approximer la récompense.

Systèmes de dialogue. L'objectif de la pro-activité dans les systèmes de dialogue est d'améliorer la convivialité des agents conversationnels en leur permettant d'initier des conversations. Alors que les systèmes de dialogue sont devenus de plus en plus populaires, les systèmes de dialogue actuels axés sur les tâches sont principalement réactifs, car les utilisateurs humains ont tendance à lancer des conversations. Les auteurs de [112] proposent d'introduire le paradigme des bandits contextuels comme cadre pour des systèmes de dialogue proactifs. Les bandits contextuels ont été le modèle de choix pour le problème de la maximisation des récompenses avec rétroaction partielle car ils correspondent bien à la description de la tâche, ils explorent également la notion de mémoire dans ce paradigme, où ils proposent deux modèles de mémoire différentiables qui agissent comme des parties du fonction d'estimation de récompense paramétrique. Le premier, les réseaux de mémoire sélective par convolution, utilise une sélection d'interactions passées dans le cadre de l'aide à la décision. Le deuxième modèle, appelé réseau de mémoire attentive contextuelle, met en oeuvre un mécanisme d'attention différentiable sur les interactions passées de l'agent. Le but est de généraliser le modèle classique des bandits contextuels aux contextes où les informations temporelles doivent être incorporées et exploitées de manière apprenable.

Systèmes de dialogue multi-domaines. Construire des agents de dialogue multi-domaines est une tâche difficile et un problème ouvert dans l'IA moderne. Dans le domaine du dialogue, la capacité d'orchestrer plusieurs agents de dialogue formés indépendamment, ou compétences, pour créer un système unifié est d'une importance particulière. Dans [113], les auteurs étudient la tâche d'orchestration du dialogue en ligne, où ils définissent l'orchestration postérieure comme la tâche de sélectionner un sous-ensemble de compétences qui répond le mieux à une entrée utilisateur en utilisant des fonctionnalités extraites à la fois de l'entrée utilisateur et

de l'individu compétences. Pour tenir compte des coûts variés associés à l'extraction des caractéristiques des compétences, ils considèrent l'orchestration postérieure en ligne avec un budget d'exécution des compétences. Ce paramètre est formalisé en tant que bandit attentif au contexte avec observations, une variante des bandits attentifs au contexte, puis l'évalue sur des ensembles de données conversationnelles simulées.

H. Détection d'une anomalie

Les auteurs de [93] étudient le problème de la détection d'anomalies dans un cadre interactif. Leur objectif est de maximiser les véritables anomalies présentées à l'expert humain après épuisement d'un budget donné. Parallèlement à cette ligne, ils formulent le problème à travers le cadre de bandit et développent un nouvel algorithme de bandit contextuel collaboratif, qui modélise explicitement les attributs et les dépendances de noeuds de manière transparente dans un cadre commun, et gère le dilemme exploration-exploitation lors de l'interrogation.

Les transactions par carte de crédit susceptibles d'être frauduleuses par les systèmes de détection automatisés sont généralement transmises à des experts humains pour vérification. Pour limiter les coûts, il est courant de ne sélectionner que les transactions les plus suspectes pour enquête. Les auteurs de [114] affirment qu'un compromis entre l'exploration et l'exploitation est impératif pour permettre l'adaptation aux changements de comportement. L'exploration consiste en la sélection et l'investigation des transactions dans le but d'améliorer les modèles prédictifs, et l'exploitation consiste à enquêter sur les transactions détectées comme suspectes. Modélisant la détection des transactions frauduleuses comme une récompense, ils utilisent un apprenant d'arbre de régression incrémentiel pour créer des grappes de transactions avec des récompenses attendues similaires. Cela permet l'utilisation d'un algorithme de bandit contextuel (CMAB) pour fournir le compromis exploration / exploitation.

I. Télécommunication

Dans [115], un modèle de bandit a été utilisé pour décrire le problème de la meilleure sélection de réseau sans fil par un dispositif multi-Radio Access Technology (multi-RAT), dans le but de maximiser la qualité perçue par l'utilisateur final. Le modèle proposé étend le modèle MAB classique de deux manières. Premièrement, il prévoit deux actions différentes: mesurer et utiliser; deuxièmement, il permet aux actions de s'étaler sur plusieurs étapes de temps. Deux nouveaux algorithmes

conçus pour tirer parti de la plus grande flexibilité offerte par le modèle muMAB ont également été introduits. Le premier, appelé mesure-utilisation-UCB1 est dérivé de l'algorithme UCB1, tandis que le second, appelé Mesure avec intervalle logarithmique, est conçu de manière appropriée pour le nouveau modèle afin de tirer parti de la nouvelle action de mesure, tout en utilisant agressivement le meilleur bras. Les auteurs de [116] démontrent la possibilité d'optimiser les performances de la technologie Long Range Wide Area Network. Les auteurs suggèrent que les nœuds utilisent des algorithmes de bandit multi-bras, pour sélectionner les paramètres de communication (facteur d'étalement et puissance d'émission). Les évaluations montrent que de telles méthodes d'apprentissage permettent de gérer bien mieux le compromis entre la consommation d'énergie et la perte de paquets qu'un algorithme Adaptive Data Rate adaptant les facteurs d'étalement et les puissances de transmission sur la base des valeurs du rapport signal sur interférence et du rapport de bruit.

J. Bandit dans les applications réelles: résumé et orientations futures

TABLE I
APPLICATION DES BANDITS DANS LA VIE RÉELLE

	MAB	Non-stat MAB	CMAB	Non-stat CMAB
Santé	✓		✓	
La finance	✓			
Tarification dynamique		✓		
Système de recommandation	✓	✓	✓	✓
Maximisation	✓			
Système de dialogue			✓	
Télécommunication	✓			
Détection d'anomalie	✓			

Le tableau I fournit un résumé des formulations de problèmes de bandit utilisées dans diverses applications spécifiques à un domaine. Le choix du modèle de bandit est souvent spécifique au domaine. Par exemple, il est évident que le bandit non stationnaire n'a pas été utilisé dans les applications de soins de santé, car des changements significatifs ne sont pas attendus dans le processus de prise des décisions de traitement, c'est-à-dire pas de transition dans l'état du patient; de telles transitions, si elles se produisaient, seraient mieux modélisées en utilisant l'apprentissage par renforcement plutôt que le bandit non stationnaire. Il existe clairement d'autres domaines où le bandit non stationnaire est un cadre plus approprié, mais il semble que ce paramètre n'ait pas encore été étudié de manière significative dans les domaines de la santé. Par exemple, la détection d'anomalie

est un domaine dans lequel un bandit contextuel non stationnaire pourrait être utilisé, car dans ce contexte, l'anomalie pourrait être contradictoire, ce qui signifie que tout bandit appliqué à ce paramètre devrait avoir une sorte de condition de dérive, afin de s'adapter à de nouveaux types d'attaques. Un autre constat est qu'aucun des travaux existants n'a tenté de développer un algorithme capable de résoudre ces différentes tâches en même temps, ou d'appliquer les connaissances obtenues dans un domaine à un autre domaine, ouvrant ainsi une direction de recherche sur le multitask et le transfer learning dans le cadre de bandit. De plus, étant donné la nature en ligne du problème de bandit, le lifelong learning serait une prochaine étape naturelle.

III. BANDIT POUR UN MEILLEUR APPRENTISSAGE AUTOMATIQUE

Dans cette section, nous décrivons comment les algorithmes de bandit pourraient être utilisés pour améliorer d'autres algorithmes, par ex. diverses techniques d'apprentissage automatique.

A. Selection d'algorithme

La sélection de l'algorithme est généralement basée sur des modèles de performances d'algorithme, appris au cours d'une séquence d'apprentissage hors ligne distincte, qui peut être d'un coût prohibitif. Dans des travaux récents, ils ont adopté une approche en ligne, dans laquelle un modèle de performance est mis à jour de manière itérative et utilisé pour guider la sélection. Le compromis exploration-exploitation qui en résultait a été représenté comme un problème de bandit avec des conseils d'experts, en utilisant un solveur existant pour ce jeu, cela nécessitait l'utilisation d'une limite arbitraire sur les temps d'exécution de l'algorithme, annulant ainsi le regret optimal du solveur. Dans [117], un cadre plus simple a été proposé pour représenter la sélection d'algorithmes comme un problème de bandit, en utilisant des informations partielles et une limite inconnue sur les pertes.

B. Optimisation des hyperparamètres

[118] a formulé l'optimisation des hyperparamètres comme un problème de bandit non stochastique d'exploration pure où des ressources prédéfinies, telles que des itérations, des échantillons de données ou des fonctionnalités sont allouées à des configurations échantillonnées aléatoirement. Ces travaux ont introduit un nouvel algorithme, Hyperband, pour ce cadre et analyse ses propriétés théoriques, offrant plusieurs garanties. En outre, Hyperband était comparé aux

méthodes d'optimisation bayésiennes populaires; il a été observé qu'Hyperband peut fournir une accélération plus grande par rapport à ses concurrents sur une variété de problèmes d'apprentissage.

C. Sélection des Features

Dans un apprentissage supervisé en ligne classique, la véritable étiquette d'un échantillon est toujours révélée au classificateur, contrairement à un bandit où une mauvaise classification se traduit par une récompense nulle, et seule la classification correcte donne la récompense. Les auteurs de [119] étudient le problème de la sélection des fonctionnalités en ligne, où le but est de faire des prédictions précises en utilisant seulement un petit nombre de fonctionnalités actives à l'aide de l'algorithme epsilon-greedy. Les auteurs de [120] abordent le problème de la sélection de fonctionnalités en ligne en abordant le problème d'optimisation combinatoire dans le cadre de bandit stochastique avec retour de bandit, en utilisant l'algorithme d'échantillonnage de Thompson.

D. Bandit pour l'apprentissage actif

L'étiquetage de tous les exemples dans un cadre de classification supervisée peut être coûteux. Les stratégies d'apprentissage actif résolvent ce problème en sélectionnant les exemples non étiquetés les plus utiles pour obtenir l'étiquette et pour former un modèle prédictif. Le choix des exemples à étiqueter peut être vu comme un dilemme entre l'exploration et l'exploitation sur l'espace d'entrée. Dans [13], une nouvelle stratégie d'apprentissage actif gère ce compromis en modélisant le problème d'apprentissage actif comme un problème de bandit contextuel. Ils proposent un algorithme séquentiel appelé Active Thompson Sampling (ATS), qui, à chaque tour, attribue une distribution d'échantillonnage sur le clusteur, échantillonne un point de cette distribution et interroge l'oracle pour cette étiquette de point d'échantillonnage. Les auteurs de [121] proposent également un algorithme d'apprentissage actif basé sur des groupes de bandits à plusieurs bras pour le problème de la classification binaire. Ils utilisent des idées telles que des limites de confiance inférieures et une régularisation auto-concordante tirée de la littérature sur les bandits à plusieurs bras pour concevoir leur algorithme.

E. Clustering

[122] considère le clustering collaboratif, qui est un paradigme d'apprentissage automatique concerné par l'analyse non supervisée de données complexes à vues

multiples à l'aide de plusieurs algorithmes fonctionnant. Les applications bien connues du clustering collaboratif incluent le clustering à vues multiples et le clustering de données distribué, où plusieurs algorithmes échangent des informations afin de s'améliorer mutuellement. L'un des principaux problèmes du clustering collaboratif et à vues multiples est d'évaluer quelles collaborations seront bénéfiques ou préjudiciables. De nombreuses solutions ont été proposées à ce problème, et toutes concluent que, à moins que deux modèles ne soient très proches, il est difficile de prédire à l'avance le résultat d'une collaboration. Pour résoudre ce problème, les auteurs de [122] proposent un algorithme collaboratif de clustering peer to peer basé sur le principe des bandits multi-bras non stochastiques pour évaluer en temps réel quels algorithmes ou vues peuvent apporter des informations utiles.

F. Apprentissage par renforcement

Les systèmes cyber-physiques autonomes jouent un rôle important dans nos vies. Pour s'assurer que les agents se comportent de manière alignée sur les valeurs des sociétés dans lesquelles ils opèrent, nous devons développer des techniques qui permettent à ces agents non seulement de maximiser leur récompense dans un environnement, mais aussi d'apprendre et de suivre les contraintes implicites assumées par la société. Dans [60], les auteurs étudient un cadre où un agent peut observer des traces de comportement de membres de la société mais n'a pas accès à l'ensemble explicite de contraintes qui donnent lieu au comportement observé. Au lieu de cela, l'apprentissage par renforcement inverse est utilisé pour apprendre de telles contraintes. C'est contraintes sont ensuite combinées avec une fonction de valeur orthogonale grâce à l'utilisation d'un orchestrateur contextuel qui choisit entre deux politiques. L'orchestrateur de bandit contextuel permet à l'agent de mélanger les politiques de manière novatrice, en prenant les meilleures actions à partir d'une politique de maximisation de la récompense.

G. Bandit pour l'apprentissage automatique: Résumé et orientations futures

Le tableau II résume les types de problèmes de bandit utilisés pour résoudre les problèmes d'apprentissage automatique mentionnés ci-dessus. Nous voyons, par exemple, que le bandit contextuel n'a pas été utilisé dans la sélection des hyperparamètres. Cette observation pourrait indiquer une direction pour les travaux futurs, où des informations secondaires pourraient être utilisées dans la sélection des caractéristiques. En outre, le bandit

TABLE II
BANDIT POUR L'APPRENTISSAGE AUTOMATIQUE

	MAB	Non Station-MAB	CMAB	Non Station-CMAB
Sélection d'algorithme		✓		
Optimisation des paramètres	✓			
Sélection des fonctionnalités	✓	✓		
Apprentissage actif	✓		✓	
Clustering	✓			
RL	✓	✓	✓	

non stationnaire a rarement été pris en compte dans ces situations problématiques, ce qui suggère également des extensions possibles des travaux actuels. Par exemple, le bandit contextuel non stationnaire pourrait être utile dans le cadre de sélection de caractéristiques non stationnaires, où trouver les bonnes caractéristiques dépend du temps et du contexte lorsque l'environnement ne cesse de changer. Notre principale observation est également que chaque technique ne résout qu'un seul problème d'apprentissage automatique à la fois; Ainsi, la question est de savoir si un paramètre de bandit et des algorithmes peuvent être développés pour résoudre simultanément plusieurs problèmes d'apprentissage automatique, et si le transfert et l'apprentissage continu peuvent être réalisés dans ce contexte. Une solution pourrait être de modéliser tous ces problèmes dans un cadre de bandit combinatoire, où l'algorithme de bandit trouverait la solution optimale pour chaque problème à chaque itération; ainsi, le bandit combinatoire pourrait en outre être utilisé comme outil pour faire progresser l'apprentissage automatique.

IV. CONCLUSIONS

Dans cet article, nous avons passé en revue certains des travaux récents les plus notables sur les applications du bandit et du bandit contextuel, à la fois dans des domaines réels et dans l'apprentissage automatique. Nous avons résumé, de manière organisée (tableaux 1 et 2), diverses applications existantes, par types de paramètres de bandit utilisés, et discuté des avantages de l'utilisation des techniques de bandit dans chaque domaine. Nous décrivons brièvement plusieurs problèmes importants et des extensions futures prometteuses. En résumé, le cadre du bandit, comprenant à la fois le bandit multi-bras et le bandit contextuel, est actuellement des domaines de recherche très actifs et prometteurs, et de multiples nouvelles techniques et applications émergent chaque année. Nous espérons que notre enquête pourra aider le lecteur à mieux comprendre certains aspects clés de ce

domaine passionnant et à avoir une meilleure perspective sur ses avancées notables et ses promesses futures.

REFERENCES

- [1] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau, "Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis," in *Machine Learning for Healthcare Conference*, pp. 67–82, 2018.
- [2] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, "Hybrid- ϵ -greedy for mobile context-aware recommender system," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 468–479, Springer, Berlin, Heidelberg, 2012.
- [3] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, "Following the user's interests in mobile context-aware recommender systems: The hybrid- ϵ -greedy algorithm," in *2012 26th International Conference on Advanced Information Networking and Applications Workshops*, pp. 657–662, IEEE, 2012.
- [4] R. Allesiardo, R. Féraud, and D. Bouneffouf, "A neural networks committee for the contextual bandit problem," in *International Conference on Neural Information Processing*, pp. 374–381, Springer, Cham, 2014.
- [5] D. Bouneffouf, "Situation-aware approach to improve context-based recommender system," *arXiv preprint arXiv:1303.0481*, 2013.
- [6] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, "Considering the high level critical situations in context-aware recommender systems," in *2nd International Workshop on Information Management for Mobile Applications*, p. 26, 2012.
- [7] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, "Exploration/exploitation trade-off in mobile context-aware recommender systems," in *Australasian Joint Conference on Artificial Intelligence*, pp. 591–601, Springer, Berlin, Heidelberg, 2012.
- [8] D. Bouneffouf, "Applying machine learning techniques to improve user acceptance on ubiquitous environment," *arXiv preprint arXiv:1301.4351*, 2013.
- [9] D. Bouneffouf, A. Bouzeghoub, and A. L. Ganarski, "Risk-aware recommender systems," in *International Conference on Neural Information Processing*, pp. 57–65, Springer, Berlin, Heidelberg, 2013.
- [10] D. Bouneffouf, "Role of temporal inference in the recognition of textual inference," *arXiv preprint arXiv:1302.5645*, 2013.
- [11] D. Bouneffouf, *DRARS, a dynamic risk-aware recommender system*. PhD thesis, 2013.
- [12] D. Bouneffouf, "Improving adaptation of ubiquitous recommender systems by using reinforcement learning and collaborative filtering," *arXiv preprint arXiv:1303.2308*, 2013.
- [13] D. Bouneffouf, R. Laroche, T. Urvoy, R. Féraud, and R. Allesiardo, "Contextual bandit for active learning: Active thompson sampling," in *International Conference on Neural Information Processing*, pp. 405–412, Springer, Cham, 2014.
- [14] D. Bouneffouf, "Towards user profile modelling in recommender system," *arXiv preprint arXiv:1305.1114*, 2013.
- [15] D. Bouneffouf, *Role de l'inference temporelle dans la reconnaissance de l'inference textuelle*. PhD thesis, Université des Sciences et de la Technologie, 2008.
- [16] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, "Contextual bandits for context-based information retrieval," in *International Conference on Neural Information Processing*, pp. 35–42, Springer, Berlin, Heidelberg, 2013.
- [17] D. Bouneffouf, "The impact of situation clustering in contextual-bandit algorithm for context-aware recommender systems," *arXiv preprint arXiv:1304.3845*, 2013.
- [18] D. Bouneffouf, "Recommandation mobile, sensible au contexte de contenus évolutifs: Contextuel- ϵ -greedy," *arXiv preprint arXiv:1402.1986*, 2014.
- [19] D. Bouneffouf and I. Birol, "Sampling with minimum sum of squared similarities for nystrom-based large scale spectral clustering," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [20] D. Bouneffouf, "Hybrid q-learning applied to ubiquitous recommender system," *arXiv preprint arXiv:1303.2651*, 2013.
- [21] D. Bouneffouf, "Mobile recommender systems methods: An overview," *arXiv preprint arXiv:1305.1745*, 2013.
- [22] D. Bouneffouf, "Exponentiated gradient exploration for active learning," *Computers*, vol. 5, no. 1, p. 1, 2016.
- [23] D. Bouneffouf, "Evolution of the user's content: An overview of the state of the art," *arXiv preprint arXiv:1305.1787*, 2013.
- [24] D. Bouneffouf, "l'apprentissage automatique", une étape importante dans l'adaptation des systèmes d'information à l'utilisateur," 2013.
- [25] D. Bouneffouf, "Context-based information retrieval in risky environment," *arXiv preprint arXiv:1409.7729*, 2014.
- [26] D. Bouneffouf, "l'étude des dimensions spécifiques du contexte dans un système de filtrage d'informations," *arXiv preprint arXiv:1405.6287*, 2014.
- [27] D. Bouneffouf, "Freshness-aware thompson sampling," in *International Conference on Neural Information Processing*, pp. 373–380, Springer, Cham, 2014.
- [28] R. Allesiardo, R. Féraud, and D. Bouneffouf, "Prise de décision contextuelle en bande organisée: Quand les bandits font un brainstorming," 2014.
- [29] D. Bouneffouf, "Temporal logic and its applications in natural language processing," 2011.
- [30] D. Bouneffouf, "R-ucb: a contextual bandit algorithm for risk-aware recommender systems," *arXiv preprint arXiv:1408.2195*, 2014.
- [31] D. Bouneffouf, "Proposition d'une technique de gestion de projet dans les startups," *arXiv preprint arXiv:1303.2317*, 2013.
- [32] D. Bouneffouf, "Learning and inference engine applied to ubiquitous recommender system,"
- [33] D. Bouneffouf, "La logique temporelle et ses applications dans le traitement du langage naturel," 2013.
- [34] D. Bouneffouf, "Ant clustering to improve situation-aware recommender systems," 2014.
- [35] D. Bouneffouf, "Contextual bandit algorithm for risk-aware recommender systems," in *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 4667–4674, IEEE, 2016.
- [36] D. Bouneffouf and R. Féraud, "Multi-armed bandit problem with known trend," *Neurocomputing*, vol. 205, pp. 16–21, 2016.
- [37] D. Bouneffouf, "Exponentiated gradient linucb for contextual multi-armed bandits," *arXiv preprint arXiv:1305.2415*, 2013.
- [38] D. Bouneffouf and I. Birol, "Theoretical analysis of the minimum sum of squared similarities sampling for nystrom-based spectral clustering," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 3856–3862, IEEE, 2016.
- [39] D. Bouneffouf, "Temporal logic in natural language processing," 2013.
- [40] D. Bouneffouf, "Optimizing an utility function for exploration/exploitation trade-off in context-aware recommender system," *arXiv preprint arXiv:1303.0485*, 2013.
- [41] D. Bouneffouf and I. Birol, "Ensemble minimum sum of squared similarities sampling for nystrom-based spectral clustering," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 3851–3855, IEEE, 2016.
- [42] D. Bouneffouf, I. Rish, G. A. Cecchi, and R. Féraud, "Context attentive bandits: contextual bandit with restricted context," in *IJCAI 2017*, 2017.
- [43] D. Bouneffouf, I. Rish, and G. A. Cecchi, "Bandit models of human behavior: Reward processing in mental disorders," in *International Conference on Artificial General Intelligence*, pp. 237–248, Springer, Cham, 2017.

- [44] D. Bouneffouf, "Drars: un système de recommandation dynamique sensible au risque,"
- [45] B. Lin, G. Cecchi, D. Bouneffouf, and I. Rish, "Adaptive representation selection in contextual bandit with unlabeled history," 2018.
- [46] D. Bouneffouf, "Nystrom sampling depends on the eigenspectrum shape of the data," 2018.
- [47] A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi, "Using contextual bandits with behavioral constraints for constrained online movie recommendation.," in *IJCAI*, pp. 5802–5804, 2018.
- [48] M. Riemer, T. Klinger, D. Bouneffouf, and M. Franceschini, "Scalable recollections for continual lifelong learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 1352–1359, 2019.
- [49] A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi, "Incorporating behavioral constraints in online ai systems," in *AAAI*, 2019.
- [50] D. Bouneffouf, "Eigenspectrum shape based nystrom sampling," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, IEEE, 2018.
- [51] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, "Following the user's interests in mobile context-aware,"
- [52] M. Riemer, M. Franceschini, D. Bouneffouf, and T. Klinger, "Generative knowledge distillation for general purpose function compression," in *NIPS 2017 Workshop on Teaching Machines, Robots, and Humans*, vol. 5, p. 30, 2017.
- [53] B. Lin, D. Bouneffouf, G. A. Cecchi, and I. Rish, "Contextual bandit with adaptive feature extraction," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 937–944, IEEE, 2018.
- [54] A. Choromanska, B. Cowen, S. Kumaravel, R. Luss, M. Rigotti, I. Rish, P. Diachille, V. Gurev, B. Kingsbury, R. Tejwani, et al., "Beyond backprop: Online alternating minimization with auxiliary variables," in *International Conference on Machine Learning*, pp. 1193–1202, 2019.
- [55] D. Bouneffouf and I. Rish, "A survey on practical applications of multi-armed and contextual bandits," in *The IEEE World Congress on Computational Intelligence (IEEE WCCI)*, 2020.
- [56] B. Djallel, A. Bouzeghoub, and A. L. Ganarski, "Risk-aware recommender systems,"
- [57] S. Upadhyay, M. Agarwal, D. Bouneffouf, and Y. Khazaeni, "A bandit approach to posterior dialog orchestration under a budget," *NIPS 2018*.
- [58] S. Liu, P. Ram, D. Bouneffouf, G. Bramble, A. R. Conn, H. Samulowitz, and A. G. Gray, "Automated machine learning via admm," *CoRR, abs/1905.00424*, 2019.
- [59] D. Bouneffouf, S. Parthasarathy, H. Samulowitz, and M. Wistub, "Optimal exploitation of clustering and history information in multi-armed bandit," *arXiv preprint arXiv:1906.03979*, 2019.
- [60] R. Noothigattu, D. Bouneffouf, N. Mattei, R. Chandra, P. Madan, K. R. Varshney, M. Campbell, M. Singh, and F. Rossi, "Interpretable multi-objective reinforcement learning through policy orchestration," *arXiv preprint arXiv:1809.08343*, 2018.
- [61] M. Yurochkin, S. Upadhyay, D. Bouneffouf, M. Agarwal, and Y. Khazaeni, "Online semi-supervised learning with bandit feedback," 2019.
- [62] B. Lin, G. Cecchi, D. Bouneffouf, J. Reinen, and I. Rish, "Reinforcement learning models of human behavior: Reward processing in mental disorders," *arXiv preprint arXiv:1906.11286*, 2019.
- [63] R. Noothigattu, D. Bouneffouf, N. Mattei, R. Chandra, P. Madan, K. R. Varshney, M. Campbell, M. Singh, and F. Rossi, "Teaching ai agents ethical values using reinforcement learning and policy orchestration," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 2–1, 2019.
- [64] C. Aggarwal, D. Bouneffouf, H. Samulowitz, B. Buesser, T. Hoang, U. Khurana, S. Liu, T. Pedapati, P. Ram, A. Rawat, et al., "How can ai automate end-to-end data science?," *arXiv preprint arXiv:1910.14436*, 2019.
- [65] A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi, "Using multi-armed bandits to learn ethical priorities for online ai systems," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 1–1, 2019.
- [66] S. Mehta, F. Rossi, K. R. Varshney, A. Balakrishnan, D. Bouneffouf, N. Mattei, R. Noothigattu, R. Chandra, P. Madan, M. Campbell, et al., "Ai ethics," 2019.
- [67] S. Liu, P. Ram, D. Vijaykeerthy, D. Bouneffouf, G. Bramble, H. Samulowitz, D. Wang, A. Conn, and A. G. Gray, "An admm based framework for automl pipeline configuration.," in *AAAI*, pp. 4892–4899, 2020.
- [68] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney, "Data augmentation for discrimination prevention and bias disambiguation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 358–364, 2020.
- [69] A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi, "Constrained decision-making and explanation of a recommendation," Jan. 16 2020. US Patent App. 16/050,176.
- [70] B. Lin, G. A. Cecchi, D. Bouneffouf, J. Reinen, and I. Rish, "A story of two streams: Reinforcement learning models for human behavior and neuropsychiatry.," in *AAMAS*, pp. 744–752, 2020.
- [71] K. Varshney, M. Campbell, M. Singh, and F. Rossi, "Teaching ai agents ethical values using reinforcement learning and policy orchestration,"
- [72] D. Bouneffouf and E. Claeys, "Hyper-parameter tuning for the contextual bandit," *arXiv preprint arXiv:2005.02209*, 2020.
- [73] B. Lin, G. Cecchi, D. Bouneffouf, J. Reinen, and I. Rish, "Unified models of human behavioral agents in bandits, contextual bandits and rl," *arXiv preprint arXiv:2005.04544*, 2020.
- [74] D. Bouneffouf and E. Claeys, "Learning exploration for contextual bandit," in *AutoML@ ICML 2019: 6th ICML Workshop on Automated Machine Learning*, 2016.
- [75] B. Lin, D. Bouneffouf, and G. Cecchi, "Online learning in iterated prisoner's dilemma to mimic human behavior," *arXiv preprint arXiv:2006.06580*, 2020.
- [76] P. Ram, S. Liu, D. Vijaykeerthy, D. Wang, D. Bouneffouf, G. Bramble, H. Samulowitz, and A. G. Gray, "Solving constrained cash problems with admm," *arXiv preprint arXiv:2006.09635*, 2020.
- [77] D. Bouneffouf, "Online learning with corrupted context: Corrupted contextual bandits," *arXiv preprint arXiv:2006.15194*, 2020.
- [78] D. Bouneffouf, S. Upadhyay, and Y. Khazaeni, "Contextual bandit with missing rewards," *arXiv preprint arXiv:2007.06368*, 2020.
- [79] D. Bouneffouf, "Location-aware approach to improve context-based recommender system," *arXiv preprint arXiv:1303.0481*, 2013.
- [80] K. Toutanova and H. Wu, "Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014.
- [81] C. S. Leung, *Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part III*. Springer, 2012.
- [82] A. T. B. Jin, *Neural Information Processing: 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014: Proceedings*. Springer, 2014.
- [83] D. Bouneffouf, "Computing the dirichlet-multinomial log-likelihood function," *arXiv preprint arXiv:2007.11967*, 2020.

- [84] D. Bouneffouf, "Spectral clustering using eigenspectrum shape based nystrom sampling," *arXiv preprint arXiv:2007.11416*, 2020.
- [85] D. Bouneffouf, S. Upadhyay, and Y. Khazaeni, "Online learning from less data: Contextual bandit with missing rewards,"
- [86] A. Gupta, Y. Ong, B. Da, L. Feng, and S. Handoko, "2016 IEEE congress on evolutionary computation (cec)," 2016.
- [87] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, "contextual-bandit algorithm for context-aware recommender system," in *International conference on neural information processing*, pp. 324–331, Springer, Berlin, Heidelberg, 2012.
- [88] D. Bouneffouf, I. Rish, and G. A. Cecchi, "Bandit models of human behavior,"
- [89] D. Bouneffouf, C. Aggarwal, H. Samulowitz, B. Buesser, T. Hoang, U. Khurana, S. Liu, T. Pedapati, P. Ram, A. Rawat, et al., "Survey on automated end-to-end data science?,"
- [90] D. Bouneffouf, "Spectral clustering using eigenspectrum shape based sampling,"
- [91] D. Bouneffouf and E. Claeys, "Online hyperparameter tuning for contextual bandits," 2018.
- [92] D. Bouneffouf, "Toward computing the dirichlet-multinomial log-likelihood function,"
- [93] K. Ding, J. Li, and H. Liu, "Interactive anomaly detection on attributed networks," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19*, (New York, NY, USA), pp. 357–365, ACM, 2019.
- [94] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.
- [95] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.
- [96] D. Bouneffouf and R. Féraud, "Multi-armed bandit problem with known trend," *Neurocomputing*, vol. 205, pp. 16–21, 2016.
- [97] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pp. 39.1–39.26, 2012.
- [98] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," in *Advances in neural information processing systems*, pp. 817–824, 2008.
- [99] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *ICML (3)*, pp. 127–135, 2013.
- [100] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," *CoRR*, 2010.
- [101] R. Allesiardo, R. Féraud, and D. Bouneffouf, "A neural networks committee for the contextual bandit problem," in *Neural Information Processing - 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014. Proceedings, Part I*, pp. 374–381, 2014.
- [102] H. Bastani and M. Bayati, "Online decision-making with high-dimensional covariates," *Available at SSRN 2661896*, 2015.
- [103] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha, "Portfolio choices with orthogonal bandit learning," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [104] X. Huo and F. Fu, "Risk-aware multi-armed bandit problem with application to portfolio selection," *Royal Society open science*, vol. 4, no. 11, p. 171377, 2017.
- [105] K. Misra, E. M. Schwartz, and J. Abernethy, "Dynamic online pricing with incomplete information using multi-armed bandit experiments," 2018.
- [106] J. Mueller, V. Syrgkanis, and M. Taddy, "Low-rank bandit methods for high-dimensional dynamic pricing," *arXiv preprint arXiv:1801.10242*, 2018.
- [107] Q. Zhou, X. Zhang, J. Xu, and B. Liang, "Large-scale bandit approaches for recommender systems," in *International Conference on Neural Information Processing*, pp. 811–821, Springer, 2017.
- [108] S. Vaswani, B. Kveton, Z. Wen, M. Ghavamzadeh, L. V. Lashmanan, and M. Schmidt, "Model-independent online learning for influence maximization," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3530–3539, JMLR. org, 2017.
- [109] Z. Wen, B. Kveton, M. Valko, and S. Vaswani, "Online influence maximization under independent cascade model with semi-bandit feedback," in *Advances in neural information processing systems*, pp. 3022–3032, 2017.
- [110] D. E. Losada, J. Parapar, and A. Barreiro, "Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems," *Information Processing & Management*, vol. 53, no. 5, pp. 1005–1025, 2017.
- [111] B. Liu, T. Yu, I. Lane, and O. J. Mengshoel, "Customized nonlinear bandits for online response selection in neural conversation models," in *AAAI, 2018*, pp. 5245–5252, 2018.
- [112] T. Silander et al., "Contextual memory bandit for pro-active dialog engagement," 2018.
- [113] S. Upadhyay, M. Agarwal, D. Bouneffouf, and Y. Khazaeni, "A bandit approach to posterior dialog orchestration under a budget," 2018.
- [114] D. J. Soemers, T. Brys, K. Driessens, M. H. Winands, and A. Nowé, "Adapting to concept drift in credit card transaction data streams using contextual bandits and decision trees," in *AAAI*, 2018.
- [115] S. Boldrini, L. De Nardis, G. Caso, M. Le, J. Fiorina, and M.-G. Di Benedetto, "mumab: A multi-armed bandit model for wireless network selection," *Algorithms*, vol. 11, no. 2, p. 13, 2018.
- [116] R. Kerkouche, R. Alami, R. Féraud, N. Varsier, and P. Maillé, "Node-based optimization of lora transmissions with multi-armed bandit algorithms," in *ICT 2018, Saint Malo, France, June 26-28, 2018*, pp. 521–526, 2018.
- [117] M. Gagliolo and J. Schmidhuber, "Algorithm selection as a bandit problem with unbounded losses," in *Learning and Intelligent Optimization, 4th International Conference, LION 4, Venice, Italy, January 18-22, 2010. Selected Papers*, pp. 82–96, 2010.
- [118] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *arXiv preprint arXiv:1603.06560*, 2016.
- [119] J. Wang, P. Zhao, S. C. Hoi, and R. Jin, "Online feature selection and its applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 698–710, 2014.
- [120] D. Bouneffouf, I. Rish, G. A. Cecchi, and R. Féraud, "Context attentive bandits: Contextual bandit with restricted context," in *IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 1468–1475, 2017.
- [121] R. Ganti and A. G. Gray, "Building bridges: Viewing active learning from the multi-armed bandit lens," *arXiv preprint arXiv:1309.6830*, 2013.
- [122] J. Sublime and S. Lefebvre, "Collaborative clustering through constrained networks using bandit optimization," in *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pp. 1–8, 2018.

This figure "fig1.png" is available in "png" format from:

<http://arxiv.org/ps/2101.00001v1>