Bandit et Semi-Bandit avec Retour Partiel : Une Stratégie d'Optimisation du Retour Utilisateur

Alexandre Letard, T Amghar, O Camp, N Gutowski

Bandit et Semi-Bandit avec Retour Partiel : Une Stratégie d'Optimisation du Retour Utilisateur

A. LETARD[1, 2], T. AMGHAR[2], O. CAMP[3], N. GUTOWSKI[3]

[1] Dpt R&D, KARA TECHNOLOGY
[2] Université d'Angers, LERIA
[3] Groupe ESEO, ERIS

alexandre.letard@kara.technology

Résumé

Aujourd'hui, dans de nombreux secteurs d'activités, les entreprises renforcent leur numérisation et proposent de nouveaux services à leurs usagers. Ces dernières années, nombre de ces services ont reposé sur des techniques d'apprentissage automatique. Pour les algorithmes de bandits-manchots combinatoires, particulièrement employés pour la recommandation, le retour utilisateur joue un rôle crucial dans l'apprentissage en ligne. Cependant, les stratégies de prise en compte de ce retour reposent essentiellement sur l'observation d'un vecteur de récompenses complet. Celui-ci reste difficile à acquérir lorsque l'utilisateur doit être directement et trop fréquemment sollicité. Dans cet article, nous proposons une nouvelle approche permettant de pallier cette problématique et maintenant une précision globale proche de celles des méthodes classiques.

Mots-clés

Apprentissage par Renforcement, Bandits-Manchots Combinatoires, Retour Utilisateur, Systèmes de Recommandation, Vitesse d'Apprentissage

Abstract

Nowadays, in most fields of activities, companies are strengthening their digitization process and offer new services to their users. In recent years, many of these services have relied on machine learning techniques. Concerning combinatorial multi-armed bandit algorithms, which are particularly employed for recommendation, user feedbacks play a crucial role for online learning. However, strategies for considering those feedbacks are essentially based on the observation of a full rewards vector which can be hard to acquire when users must be directly and too frequently solicited. Herein, we propose a novel approach which overcomes these limitations, while providing a level of global accuracy similar to that obtained by classical competitive methods.

Keywords

Reinforcement Learning, Combinatorial Multi-Armed Bandits, User Feedbacks, Recommendations System, Learning Speed

1 Introduction

De nos jours, les systèmes de recommandations basés sur des méthodes d'apprentissage automatique sont devenus courants dans de nombreux domaines [12]. Parmi les techniques employées, celles reposant sur les bandits-manchots obtiennent de bons résultats en matière de précision globale [12, 6]. C'est le cas, plus particulièrement des bandits-manchots combinatoires [17]. Dans un cadre industriel, certains secteurs d'activités, comme la navigation de plaisance [7], amorcent une transition numérique afin de proposer des services similaires à leurs usagers. Dans le domaine de l'habitat mobile, auquel appartient la navigation de plaisance [7], on observe une concentration des problématiques inhérentes à un lieu de vie et à un véhicule. Il existe ainsi de multiples modes d'utilisation d'un véhicule habitable qui peut être une résidence, principale ou secondaire, ou un moyen de transport, ou encore un outil de dépassement de soi. Ces modes d'utilisation sont propres à chaque usager et à chaque contexte d'usage. Au cours des dernières années, des travaux ont été entrepris pour favoriser la transition numérique du nautisme [14, 22]. Ces travaux portent essentiellement sur l'automatisation des manoeuvres de navigation et ne traitent pas des autres aspects de l'habitat mobile. Or, afin d'apporter des recommandations pertinentes aux navigateurs et ainsi améliorer leurs expériences maritimes, il est nécessaire de prendre en considération l'ensemble de ces aspects. À terme, notre objectif est la mise en oeuvre d'un bateau intelligent, EVA [1], dont le comportement sera guidé par les besoins des utilisateurs, en regard de leur mode d'utilisation du bateau. Nous viserons à réduire les risques de pénuries d'énergie en mer et les impacts environnementaux en optimisant l'utilisation du bateau. Nous mettrons en oeuvre des techniques de bandits-manchots combinatoires afin d'identifier les fonctionnalités "juste nécessaires" parmi celles disponibles à bord, pour satisfaire les usagers.

Afin d'effectuer des recommandations personnalisées, les techniques de bandits-manchots combinatoires considèrent le retour utilisateur exprimé à chaque recommandation [10]. Dans la littérature, les stratégies de prise en compte

1. Entité de Voyages Automatisée

du retour utilisateur les plus fréquemment exploitées reposent sur un vecteur de récompenses complet pour l'ensemble des recommandations effectuées [20, 10]. Ce vecteur peut se révéler difficile à acquérir, notamment lorsque les récompenses dépendent d'un retour explicite de l'utilisateur, p.ex sous la forme d'un score ou d'une évaluation. À ce titre, dans cet article nous expérimentons et évaluons une autre méthode permettant de favoriser l'application des techniques de bandits-manchots combinatoires au sein de systèmes interactifs, en vue de leur intégration dans des secteurs d'activités nouvellement connectés.

Nous proposons une approche appliquant à la fois des considérations de type bandit et semi-bandit sur un sous-ensemble d'éléments recommandés de taille variable. Nous nommons cette méthode "Partial Bandit with Semi-Bandit" (*P-BSB*). Nous proposons trois stratégies pour déterminer le sous-ensemble à observer : *Reinforce - RE*, *Optimal Exploration - OE* et *Randomized - RD*.

Au travers de nos expérimentations, nous appliquons une approche combinatoire à plusieurs algorithmes de bandits manchots à tirages simples. Nous observons que cette stratégie permet l'emploi des techniques de bandits-manchots combinatoires avec un retour utilisateur restreint. Aussi, nous constatons que les résultats de précision globale et d'itération de convergence obtenus pour un horizon supérieur à 10 000 itérations sont proches de ceux obtenus avec des méthodes classiques de bandit et semi-bandit.

En résumé, nos contributions visent à permettre : 1) l'application d'une méthode proposant un tirage multiple à des algorithmes de bandits manchots de l'état de l'art couramment utilisés en tirage simple, notamment *UCB2* [4], qui, à notre connaissance, n'a pas été employé dans un cadre combinatoire ; 2) l'évaluation et la comparaison de ces algorithmes en termes de précision globale et d'itération de convergence sur plusieurs jeux de données issus d'applications réelles ; 3) la formalisation et la proposition d'une nouvelle méthode de considération du retour utilisateur – "Partial Bandit with Semi-Bandit" – pour les algorithmes de bandits manchots que nous déclinons selon trois variantes, afin de réduire les contraintes liées à l'acquisition du retour utilisateur.

Cet article est organisé comme suit : la section 2 expose les notions fondamentales inhérentes aux techniques employées, quelques travaux connexes de la littérature, et nos motivations. La section 3 formalise notre problématique et l'approche que nous proposons pour y répondre. La section 4 analyse les résultats d'expérimentations effectuées avec des jeux de données issus d'applications réelles. Enfin, dans la section 5, nous concluons et exposons les futurs travaux envisagés.

2 Travaux Connexes et Motivations

Cet article traite des bandits-manchots combinatoires et notamment des stratégies de prise en compte du retour utilisateur dans le cadre de leur apprentissage. Ainsi, dans cette section, avant d'aborder les spécificités des bandits-manchots combinatoires [9] (*Combinatorial Multi-Armed Bandit* : *COM-MAB* et *Combinatorial Contextual Multi-Armed Bandit* : *COM-CMAB*) et des approches couramment employées pour exploiter le retour de l'utilisateur, nous rappellerons le problème du bandit-manchot [19] et le problème des bandits-manchots contextuels [15]. Enfin, nous évoquerons les travaux connexes de la littérature et présenterons nos motivations à envisager la combinaison de deux prises en compte partielles du retour utilisateur au sein d'applications réelles de bandits manchots combinatoires.

2.1 Les bandits manchots

Il existe une vaste littérature sur le problème de bandits-manchots (*MAB*) largement étudiés depuis leur première formulation par Robbins en 1952 [19]. Il en résulte aujourd'hui de nombreuses approches [12] : stochastiques [3], non stochastiques [5] ou bayésiennes [1]. Un problème de bandits-manchots est composé d'un ensemble $\mathcal{A} = \{a_1, ...a_m\}$ de *m* bras indépendants, où chaque bras $a \in \mathcal{A}$ correspond à un élément à recommander. Dans le cadre des systèmes de recommandation, à chaque itération $t \in [1,\ T]$, *T* étant l'horizon, un agent sélectionne, suivant une politique π, un bras $a_t \in \mathcal{A}$, correspondant à un élément à recommander et le propose à l'utilisateur. Une partie du vecteur de récompenses Y_t [2] associé à $\mathcal{A}$ est dévoilée à l'agent qui perçoit alors une unique récompense $r_{t,a}$ pour l'élément recommandé. Dans cet article, nous nous intéresserons plus particulièrement au problème de *MAB* de type Bernoulli où $r_{t,a} \in \{0, 1\}$ avec $r_{t,a} = 0$ si l'utilisateur ne valide pas la recommandation qui lui a été faite et $r_{t,a} = 1$ s'il la valide [12]. Dans le cadre d'une approche stochastique, où les récompenses sont considérées comme étant des variables indépendantes et identiquement distribuées entre les bras, l'objectif d'un algorithme de *MAB* est de minimiser le regret $\rho_T = T\mu^* - \sum_{t=1}^{T} r_{t,a}$, où μ^* désigne l'espérance de récompense associée au bras optimal, sans connaissance a priori de la distribution des probabilités de récompenses $\mu_a \in [0, 1]$ associées à chacun des bras a de $\mathcal{A}$. Chercher à minimiser le regret observé revient à maximiser la précision globale $Acc(T) = \frac{\sum_{t=1}^{T} r_{t,a}}{T}$, qui est également couramment exploitée comme critère d'évaluation au sein de la littérature [12].

Dans le cadre particulier des bandits-manchots contextuels (*CMAB*), le contexte de l'utilisateur doit être pris en considération. Il est traduit sous la forme d'un vecteur $x \in \mathcal{X}$, $x \subseteq \mathbb{R}^d$ encodant les *d* caractéristiques de l'utilisateur et de l'environnement dans lequel il évolue, p.ex., le profil (âge, sexe, métier), les préférences, l'environnement (localisation, quartier), ou encore l'activité qu'il réalise. Dans cette variante contextuelle, on considère qu'il existe une dépendance entre l'espérance de récompense d'un bras et le contexte observé. Dans les cas d'une dépendance linéaire, l'espérance s'exprime en fonction du contexte comme suit : $\mathbb{E}[r_{t,a}|x_t] = \hat{\theta}_{t,a}^{\top} x_t$, où $\hat{\theta}_{t,a}$ est un vecteur de coefficients, associé au bras *a*, initialement nul et estimé à l'itération *t*.

2. Y_t est supposé existant mais ne peut être, en réalité, observé qu'en partie ($r_{t,a}$ en cas de tirage simple, R_t dans un cadre combinatoire).

2.2 Les bandits-manchots combinatoires

Les problèmes de bandits manchots combinatoires correspondent à une généralisation des problèmes de *MAB* et *CMAB* où l'utilisateur se voit proposer un super-bras, S_t, constitué de k éléments, tels que $S_t = \cup_{i=1}^{k} a_i$, avec $a_i = argmax^{\pi}_{a \in \mathcal{A} \setminus \{a_1,..,a_{i-1}\}} \mathbb{E}[R_{t,a}|x_t]$. À notre connaissance,la principale approche combinatoire est le tirage multiple. Cette méthode constitue S_t dynamiquement en répétant l'action de recommandation d'une ou plusieurs instances à tirages simples pour sélectionner k bras [9, 17, 10] ensuite agrégés en S_t. Ainsi, l'apprentissage s'effectue toujours au travers des bras individuels $a \in \mathcal{A}$.

Ainsi, la valeur de récompense associée à la recommandation S_t, utilisée dans l'apprentissage de l'agent, est exprimée par $S_t^\top R_t = \sum_{i=1}^{k} S_{t,i}\, R_{t,i}$, où R_t est le vecteur de récompense observé, de dimension *k*. Nous nommons ϕ la stratégie de prise en compte du retour utilisateur déterminant : 1) la construction de R_t à partir de Y_t et S_t ; 2) la politique d'apprentissage de l'agent.

Cette évolution des techniques *MAB*/*CMAB* a été employée dans de nombreux secteurs d'activités tels que les systèmes de recommandation, la finance ou le domaine médical [6].

Ainsi, l'algorithme 1 [9, 17, 10] est utilisé dans les expérimentations présentées ici. Dans le cadre de cet article, une valeur de récompense globale, $r_t \in \{0, 1\}$, où $r_t = 1$ si au moins un des éléments de S_t est validé par l'utilisateur, 0 sinon, est employée pour le calcul de la précision globale de l'algorithme tel que défini dans la sous-section 2.1.

2.3 Prise en compte du retour utilisateur

Différentes stratégies de prise en compte du retour utilisateur ont été développées pour les bandits-manchots combinatoires. Ces variantes peuvent être majoritairement regroupées en deux approches principales : **bandit** [13] et **semi-bandit** [20, 10].

Dans l'approche **bandit**, l'agent observe uniquement une récompense cumulée pour le "super-bras" S_t, sans connaître la valeur de retour propre à chacun des *k* bras le constituant : $R_{t,\,\phi_B} = S_t^\top R_t$. L'approche **semi-bandit** permet l'observation de la récompense spécifique de chaque bras $a_{t,i}$ constituant S_t : $R_{t,\,\phi_{SB}} = \cup_i\, S_{t,i}\, R_{t,i}$. Dans les deux cas, l'ensemble du vecteur de récompense R_t, de dimension *k*, reste nécessaire à l'apprentissage.

Dans de récents travaux, il est remarqué que l'approche de type **semi-bandit** est prépondérante [20]. Il en existe également de nombreuses déclinaisons permettant son exploitation dans certains cadres applicatifs, p.ex., le modèle en cascade où le retour utilisateur est exprimé par un clic sur une recommandation et où la position de l'élément cliqué est exploitée pour déterminer implicitement les autres valeurs de retours [16].

2.4 Retours partiels

Une stratégie *partielle* de prise en compte des retours utilisateur correspond à une approche où le vecteur de récompenses observées R_t est seulement défini sur une partie $P_t \subseteq S_t$ de l'ensemble des éléments recommandés à l'utilisateur. Formellement, lors d'une considération non-partielle, la dimension du vecteur de récompenses observées s'exprime par $|R_t| = |S_t|$, tandis que pour les approches partielles elle s'exprime $|R_t| = l$ avec $l < |S_t|$. Parmi les exemples de la littérature [11, 18], l est considéré comme une constante. Ainsi, Grant et al. [11] emploie une approche de semi-bandit partiel reposant sur un filtrage par application d'une loi binomiale. Luedtke et al. [18] exploite aussi une approche de type semi-bandit partiel où un sous-ensemble de S_t, est sélectionné uniformément parmi tous les sous-ensembles de S_t de cardinalité l.

L'approche **P-BSB** se différencie par la détermination d'un sous-ensemble de taille variable. L'objectif applicatif de cette nuance est de permettre, à chaque itération, l'exploitation du nombre maximal de retours que l'usager est prêt à prodiguer sans l'excéder. Une autre distinction porte sur la stratégie employée pour construire R_t. Avec **P-BSB**, l'identification des retours à solliciter auprès de l'utilisateur est soit aléatoire (variante **RD**), soit basée sur l'apprentissage réalisé jusqu'à l'itération $t-1$ (variantes **RE** et **OE**). Enfin, notre méthode diffère par l'observation d'une récompense double (R_{t,ϕ_B} et $R_{t,\phi_{SB}}$) lorsqu'un bras a effectivement apporté satisfaction à l'utilisateur.

Algorithme 1 : Bandit à tirages multiples

Entrées : π : Une instance d'une politique de bandit à tirages simples et ses paramètres particuliers.
$\mathcal{A}$: L'ensemble des bras disponibles.
k : Le nombre d'éléments à recommander à chaque itération.
Y_t, Le vecteur de récompenses réelles.
T : L'horizon.
$x \in \mathcal{X}$: Le contexte utilisateur.
$\phi(S_t, Y_t)$: La stratégie de considération du retour utilisateur.

Initialisation : Initialiser l'instance conformément aux besoins de π

pour $t = 1$ *à* T **faire**
  Considérer $x_t \in X$: un utilisateur u et son contexte
  $S_t \leftarrow \emptyset$
  pour $i = 1$ *à* k **faire**
    Sélectionner l'élément $a_i \in \mathcal{A} \setminus \{S_t\}$ selon x_t et π
    $S_t \leftarrow S_t \cup a_i$
  fin
  Recommander S_t à l'utilisateur u
  Recevoir la récompense globale r_t de la recommandation $S_t, r_t \in \{0, 1\}$
  Déterminer R_t à partir de Y_t et S_t selon ϕ
  Mettre à jour la politique π avec R_t selon π et ϕ
fin

2.5 Motivations

Les algorithmes de *MAB* ou de *CMAB* visent à maximiser leur précision globale [12]. À cette fin, la prise en compte du retour utilisateur joue un rôle majeur. Cependant, l'acquisition d'un vecteur de récompense R_t complet, nécessaire dans les considérations bandit et semi-bandit, peut s'avérer difficile voire impossible en situation réelle.

De nombreux secteurs d'activités, comme la navigation de plaisance, amorcent leur transition numérique et ne disposent donc aujourd'hui d'aucun jeu de données permettant l'entraînement hors ligne d'un agent. Dans ce cadre applicatif, les usagers risquent de se détourner d'une application s'ils ne sont pas satisfaits des recommandations proposées. Il est alors crucial pour le système de recommandation d'acquérir rapidement en ligne une connaissance suffisante. Ainsi, nous soutenons que l'itération de convergence de la précision globale, indicatrice de la vitesse d'apprentissage de l'agent, doit être prise en considération comme un critère d'évaluation à part entière des algorithmes étudiés.

Par ailleurs, des approches telles que celles en cascade [16] restent délicates à employer si le retour utilisateur doit être explicitement sollicité, p.ex., sous la forme d'un score pour déterminer les points de valeur parmi les étapes et activités - qui seraient les bras disponibles d'un algorithme de bandit manchot combinatoire - d'un voyage défini et recommandé par l'agent. Le nombre potentiellement important de retours demandés auprès de l'usager pour satisfaire une telle configuration pourrait en effet le détourner de la solution.

Ainsi, l'approche *Partial Bandit with Semi-Bandit* (*P-BSB*), proposée dans cet article, repose sur l'identification d'un sous-ensemble de R_t, $\mathcal{P}_t$, de cardinalité ρ variable, correspondant au nombre de sollicitations auxquelles l'usager accepte de répondre. **P-BSB** emploie ensuite une approche de type bandit sur S_t et une approche de type semi-bandit sur P_t. Cette double attribution de récompense à certains bras vise à accroître la vitesse d'apprentissage en avantageant les bras pour lesquels une satisfaction de l'utilisateur est effectivement observée. Au travers de cette approche, l'objectif est de faciliter l'utilisation des techniques de bandits-manchots combinatoires sur un plus large spectre d'applications du monde réel en : 1) réduisant les sollicitations auprès des utilisateurs ; 2) conservant des performances similaires à celles observées avec les méthodes classiques.

3 Définition du problème et méthodes

Dans cette section, nous formulons notre problème et décrivons notre nouvelle méthode. Celle-ci porte sur la prise en compte du retour utilisateur et repose sur la combinaison des stratégies de "bandit" et "semi-bandit" couramment employées dans la littérature. Nous les appliquons avec un nombre restreint de retours utilisateur observés.

3.1 Énoncé du problème

Soit $\mathcal{X} \subseteq \{0,1\}^d$ l'ensemble des vecteurs de contexte de dimension d caractérisant un utilisateur et son environnement p.ex., $x \in \mathcal{X}$ est un vecteur binaire encodant les caractéristiques des utilisateurs demandant une recommandation p.ex., des activités à réaliser au cours d'un voyage, à l'instant *t* d'un horizon fini *T* déterminé à l'avance. Dans le cas non contextuel c.-à-d., en l'absence de contexte ou sans prise en compte du contexte par l'algorithme, alors $\forall t \in T, x_t = \vec{0}$, le vecteur x_t est alors limité à un simple identifiant.
Soit $\mathcal{A} = \{a_1, ..., a_m\}$ l'ensemble des éléments pouvant être recommandé par un algorithme de *MAB* ou *CMAB* donné de politique π et $\mu = \{\mu_1, ..., \mu_m\}$ la distribution des espérances de récompenses associées à chaque bras *a* de $\mathcal{A}$ selon π. Soit S_t le sous-ensemble constitué d'éléments de $\mathcal{A}$, de dimension $k < m$ à l'itération $t \in [1, T]$. À chaque pas de temps t on recommande un super-bras S_t, déterminé selon π et μ_t, à un utilisateur u_t se présentant avec son vecteur de contexte x_t. Enfin, soient $r_t \in \{0,1\}$ la récompense globale associée à S_t utilisée pour le calcul de la précision globale, $Acc^{\pi}(T)$ de l'agent, Y_t un vecteur associant une récompense réelle à chacun des bras $\{a_1, ..., a_m\}$ de $\mathcal{A}$ et $R_t \subseteq Y_t$ le vecteur de récompense exprimé par l'usager et effectivement observé par l'agent. Dans les cas des stratégies de type bandit et semi-bandit il est supposé que $R_t = Y_t$ pour les *k* bras de S_t. Or, dans plusieurs applications du monde réel, lorsque les récompenses observées ne peuvent être obtenues que par une sollicitation explicite de l'usager pour l'ensemble des bras constituant S_t, cette nécessité devient alors très difficile à satisfaire. Ce constat peut se révéler critique pour nombre d'applications du monde réel. Ainsi, une nouvelle approche visant à réduire les sollicitations des utilisateurs tout en maintenant un apprentissage efficace semble nécessaire. La sous-section suivante présente une nouvelle méthode exploitant une combinaison des stratégies de type bandit et semi-bandit sur un sous-ensemble restreint de S_t.

3.2 Partial Bandit with Semi-Bandit : P-BSB

Soit un cadre applicatif où le retour utilisateur doit être explicitement demandé à l'usager et où $|R_t| = \rho \leq k$ est donc une variable aléatoire représentative de la capacité de l'utilisateur à effectuer un retour à l'agent, pouvant dépendre p.ex., de sa disponibilité, de son intérêt à répondre, de son humeur. Dans cet article, cette capacité est désignée sous le terme de "patience" de l'utilisateur. L'approche *P-BSB* vise à construire R_t et à déterminer son utilisation pour l'apprentissage de l'agent. Cette méthode correspond à une application des lignes 10 et 11 de l'algorithme 1. La première étape de l'approche *P-BSB* consiste à identifier un sous-ensemble $\mathcal{P}_t \subseteq \mathcal{S}_t$ de cardinalité ρ pour lequel des récompenses pourront effectivement être observées par l'agent, tel que $P_t = \cup_{i=1}^{\rho} a_i$. Pour ce faire, *P-BSB* est décliné selon trois variantes pour déterminer les bras a_i considérés :

— **Reinforce - RE** : sélectionne les ρ bras de S_t

ayant l'espérance de récompense $\mathbb{E}[R_{t,a}|x_t]$ la plus haute, c.-à-d. :

$$a_i = argmax^{\pi}_{a \in S_t \setminus \{a_1,..,a_{i-1}\}} \mathbb{E}[R_{t,a}|x_t] \quad (1)$$

— **Optimal-Exploration - OE** : sélectionne les ρ bras de S_t ayant été le moins fréquemment observés à l'itération *t*, c.-à-d. :

$$a_i = argmin_{a \in S_t \setminus \{a_1,..,a_{i-1}\}} obs_{a,t} \quad (2)$$

où $obs_{a,t}$ est le nombre de fois qu'une récompense a été observée pour le bras a à l'itération t.

— **Randomized - RD** : sélectionne aléatoirement ρ bras distincts de S_t, c.-à-d. :

$$a_i = random(S_t) \quad (3)$$

où $random(S_t)$ correspond à la sélection aléatoire d'un bras dans S_t n'ayant pas été précédemment choisi.

La seconde étape de *P-BSB* est commune à toutes ces variantes et consiste à acquérir le vecteur R_t de récompenses des ρ bras considérés à partir de Y_t, ou autrement dit, de l'utilisateur :

$$R_t = \cup_{i \in P_t} Y_{t,i}$$

Enfin, la troisième étape consiste à appliquer une stratégie de type bandit sur l'ensemble des *k* bras de S_t et une stratégie de type semi-bandit sur les ρ bras de P_t à partir de l'échantillon R_t observé [3] :

$$\forall a \in S_t, r_{t,a} = r_{t-1,a} + R_{t,B}$$

et si $a \in P_t$ alors $r_{t,a} = r_{t,a} + R_{t,SB_a}$

Où $r_{t,a}$ désigne l'ensemble des récompenses perçues pour le bras *a* à l'itération *t*, et avec :

$$R_{t,B} = \mathcal{P}_t^{\top} \, R_t$$

et $\forall i \in P_t$:

$$R_{t,SB} = \cup_i \, P_{t,i} \, R_{t,i}$$

Cet article détaille la variante **RE** dans l'Algorithme 2. Les autres variantes suivent la même trame globale et ne diffèrent que par leur stratégie de constitution de $\mathcal{P}_t$. Ainsi, pour employer ces variantes, il convient de remplacer la ligne 2 de l'Algorithme 2 par les éléments de l'Équation 2 pour **OE** et de l'Équation 3 pour **RD**.
L'objectif de **RE** est de favoriser une exploitabilité applicative rapide en favorisant une action optimiste de l'agent. **OE**, quant à lui, suit un objectif exploratoire et cherche donc à renforcer la connaissance de l'agent sur la distribution des espérances de récompenses $\mu_1, ..., \mu_k$ des bras recommandés. Enfin, **RD** applique une stratégie de sélection aléatoire permettant un comportement plus proche d'une prise en compte semi-bandit classique.

3. Lorsque $a \in \mathcal{P}_t$, l'agent observe donc deux récompenses pour le bras a : $R_{t,B}$ et R_{t,SB_a}. Si $\rho = 0$, l'agent n'observe alors aucune récompense à l'itération *t*.

4 Expérimentations et résultats

Nous présentons dans cette section l'évaluation empirique hors ligne de notre méthode. Cette phase d'expérimentation est préliminaire à l'intégration en ligne de notre méthode dans notre système de recommandation en environnement marin.
Ainsi dans cette section, nous commençons par présenter les jeux de données et les algorithmes employés pour évaluer notre approche. Nous exposons ensuite le protocole expérimental. Enfin nous présentons et analysons les résultats obtenus.

Algorithme 2 : *P-BSB* - RE

Entrées : S_t, Le super-bras recommandé à l'utilisateur.
Y_t, Le vecteur de récompenses réelles.
ρ_t, Le nombre de bras de S_t dont les récompenses peuvent être observées.

1 **tant que** $|\mathcal{P}_t| < \rho_t$ **faire**
2 Constituer $\mathcal{P}_t$ tel que $\mathcal{P}_t = \cup_i \, argmax^{\pi}_{a \in S_t \setminus \{a_1,..,a_{i-1}\}} \mathbb{E}[R_{t,a}|x_t]$ (selon l'Équation 1)
3 **fin**
4 **pour** $i \in P_t$ **faire**
5 Construire R_t tel que $R_t = \cup_i \, Y_{t,i}$
6 Appliquer la stratégie semi-bandit à R_t : $R_{t,SB} = \cup_i \, P_{t,i} \, R_{t,i}$
7 **fin**
8 Appliquer la stratégie bandit à R_t : $R_{t,B} = \mathcal{P}_t^{\top} \, R_t$
9 **pour** $a \in S_t$ **faire**
10 Mettre à jour la politique π avec $r_{t,a} = r_{t-1,a} + R_{t,B}$
11 **si** $a \in \mathcal{P}_t$ **alors**
12 Mettre à jour la politique π avec $r_{t,a} = r_{t,a} + R_{t,SB_a}$
13 **fin**
14 **fin**

4.1 Jeux de données

Nous évaluons notre approche sur cinq jeux de données issus d'applications réelles (cf. Tableau 1) :

— **Covertype** [4] ainsi que **Poker Hand** [5] offrent un nombre important de contextes utilisateur et permettent ainsi de passer à l'échelle ;

4. https ://archive.ics.uci.edu/ml/datasets/covertype
5. https ://archive.ics.uci.edu/ml/datasets/Poker+Hand
6. https ://www.kaggle.com/vikashrajluhaniwal/jester-17m-jokes-ratings-dataset
7. https ://www.kaggle.com/assopavic/recommendation-system-for-angers-smart-city
8. https ://grouplens.org/datasets/movielens/100k/

Jeu de données	Instances	Bras	Variables contextuelles
Covertype [4]	581 012	7	54
Poker [5]	1 025 010	10	10
RS-ASM [7]	2 152	18	50
Jester [6]	59 132	150	0
MovieLens [8]	942	1682	23

Tableau 1 – Jeux de données employés dans nos expérimentations

— **RS-ASM** [7] est un jeu de données pour la recommandation de services dans la ville intelligente [12] ;
— **Jester** [6] est un jeu de données pour la recommandation de blague où aucune information de contexte n'est disponible ;
— **MoviesLens** [8] est un jeu de données spécifique pour la recommandation de films.

Jester et **MoviesLens** représentent les cas où le nombre de bras disponibles est important et où le retour utilisateur est exprimé sous la forme d'un score allant de 0 à 5. Pour cette expérience, nous définissons un seuil $s = 4$ où $R_{t,a} = 1$ si le score est supérieur ou égal à s, 0 sinon.

4.2 Algorithmes

La méthode évaluée porte sur la prise en compte du retour utilisateur. Ainsi, elle fonctionne indépendamment de la politique π suivie par l'agent c.-à-d., indépendamment de l'algorithme de *COM-MAB* / *COM-CMAB* employé. Nous mettons en oeuvre l'algorithme 1 avec plusieurs algorithmes à tirages simples connus de la littérature avant de les évaluer en termes de précision globale et d'itération de convergence.
Ainsi, dans cet article et à la lumière des précédentes évaluations effectuées dans la littérature sur les bandits-combinatoires [9, 17, 8], nous considérons les algorithmes suivants :

— **MAB** : *ε-greedy* [21] avec $\varepsilon = 0.0009$, *Thompson Sampling (TS)* [1], *UCB* [3] et *UCB2* [4] ;
— **CMAB** : *LinUCB* [15] et *LinTS* [2].

4.3 Protocole expérimental

Au cours de nos expériences et pour chaque algorithme, afin de simuler un flux de données d'utilisateurs se présentant dans un contexte donné pour recevoir une recommandation (voir ligne 2 de l'algorithme 1), une sélection aléatoire est réalisée séquentiellement parmi les contextes disponibles dans l'ensemble du jeu de données jusqu'à un horizon fixe $T = 10000$. L'itération de convergence t_c considérée dans cet article correspond à la première itération *t* à partir de laquelle la précision globale demeure équivalente à la précision globale finale $Acc(T)$ (voir calcul à la Sous-Section 2.1), à $\delta = 1\%$ près :

$\forall t \geq t_c$:

$$Acc(T) - \delta \leq Acc(t) \leq Acc(T) + \delta, \text{ avec } \delta = 0.01$$

Algorithme	**Stratégie**	$Acc(T)$	t_c
ε-greedy	Bandit	$0,833 \pm_{0,002}$	$1461 \pm_{1746}$
	Semi-Bandit	$0,859 \pm_{0,004}$	$840 \pm_{912}$
	P-BSB-RE	$0,840 \pm_{0,005}$	$2476 \pm_{1359}$
	P-BSB-OE	$0,836 \pm_{0,004}$	$411 \pm_{265}$
	P-BSB-RD	$0,838 \pm_{0,002}$	$1288 \pm_{1113}$
TS	Bandit	$0,825 \pm_{0,002}$	$928 \pm_{1472}$
	Semi-Bandit	$0,857 \pm_{0,003}$	$1938 \pm_{1511}$
	P-BSB-RE	$0,845 \pm_{0,004}$	$1206 \pm_{1211}$
	P-BSB-OE	$0,837 \pm_{0,003}$	$545 \pm_{541}$
	P-BSB-RD	$0,839 \pm_{0,007}$	$1079 \pm_{1396}$
UCB	Bandit	$0,832 \pm_{0,005}$	$1171 \pm_{1304}$
	Semi-Bandit	$0,842 \pm_{0,002}$	$4163 \pm_{1512}$
	P-BSB-RE	$0,830 \pm_{0,004}$	$1530 \pm_{1566}$
	P-BSB-OE	$0,823 \pm_{0,004}$	$774 \pm_{905}$
	P-BSB-RD	$0,826 \pm_{0,002}$	$1580 \pm_{1542}$
UCB2	Bandit	$0,796 \pm_{0,002}$	$948 \pm_{1041}$
	Semi-Bandit	$0,796 \pm_{0,002}$	$886 \pm_{950}$
	P-BSB-RE	$0,790 \pm_{0,002}$	$1554 \pm_{1611}$
	P-BSB-OE	$0,801 \pm_{0,003}$	$889 \pm_{1213}$
	P-BSB-RD	$0,792 \pm_{0,001}$	$1734 \pm_{2235}$

Tableau 2 – Résultats pour une application non-contextuelle avec $k = 10$ sur le jeu de données **Jester**.

Chacun des algorithmes *COM-MAB / COM-CMAB* est employé sur les cinq jeux de données pour réaliser des recommandations de 3 éléments ($k = 3$). Ces expériences sont réalisées en employant successivement les stratégies de considération du retour utilisateur **bandit**, **semi-bandit**, **RE**, **OE** et **RD**, afin de permettre une comparaison des approches.
Lorsque l'une des variantes de **P-BSB** est appliquée, la "patience" ρ de l'usager est simulée par une variable aléatoire comprise entre 0 et *k*, générée à chaque itération.
Le même procédé est employé pour effectuer des recommandations à 10 éléments ($k = 10$), en faisant varier ρ entre 0 et 4, sur les jeux de données **Jester** et **MoviesLens**, disposant d'un nombre de bras important, afin d'expérimenter des situations où $\rho \ll k \ll m$.
Ainsi, pour chacun des différents cas et pour chaque approche, 10 expériences de 10 000 itérations sont simulées.
Les tableaux 2 et 3 présentent les moyennes et écarts-type obtenus pour les critères de précision globale et d'itération de convergence dans les expérimentations où $k = 10$ avec $0 \leq \rho \leq 4$. Ce cas est particulièrement intéressant dans la mesure où le nombre de bras pour lesquels l'agent ne pourra pas observer de récompense à l'itération *t* est plus important, l'expérience est donc plus représentative des résultats pouvant être attendus pour l'application visée à terme.
À la sous-section suivante nous nous focaliserons sur l'interprétation et l'analyse de ces résultats et indiquerons si les tendances observées sont vérifiées au travers de nos autres expérimentations.

4.4 Analyse des résultats

Afin d'observer l'impact d'une approche dans l'apprentissage d'un agent dans un cadre applicatif spécifique indépendamment de l'algorithme *COM-MAB / COM-CMAB*

Algorithme	Stratégie	$Acc(T)$	t_c
LinTS	Bandit	$0,996 \pm 0,001$	330 ± 253
	Semi-Bandit	$0,995 \pm 0,001$	732 ± 472
	P-BSB-RE	$0,989 \pm 0,001$	627 ± 505
	P-BSB-OE	$0,986 \pm 0,001$	391 ± 417
	P-BSB-RD	$0,989 \pm 0,001$	491 ± 330
LinUCB	Bandit	$0,994 \pm 0,001$	944 ± 513
	Semi-Bandit	$0,992 \pm 0,001$	1582 ± 596
	P-BSB-RE	$0,982 \pm 0,001$	1034 ± 541
	P-BSB-OE	$0,979 \pm 0,001$	644 ± 399
	P-BSB-RD	$0,982 \pm 0,002$	1028 ± 662

Tableau 3 – Résultats pour une application contextuelle avec $k = 10$ sur le jeu de données **MoviesLens**.

choisi, nous considérons la moyenne des résultats obtenus par les algorithmes employés. Ainsi, les approches sont comparées à partir des résultats obtenus avec l'équation suivante :

$$\forall \pi \in \Pi : M_\phi = \sum_{\pi=1}^{|\Pi|} \frac{Acc_\phi^\pi(T)}{|\Pi|} \quad (4)$$

Ce procédé est également exploité pour les comparaisons d'itération de convergence : il suffit de remplacer les valeurs $Acc(T)$ par les valeurs correspondantes des colonnes t_c des tableaux de résultats 2 et 3.

4.4.1 Observations spécifiques - Jester et MoviesLens

Le tableau 4 présente les résultats de précision globale - $Acc(T)$ - et d'itération de convergence - t_c - observés en moyenne (selon l'équation 4) pour chacune des approches considérées sur les jeux de données Jester et MoviesLens :

	Jester		**MovieLens**	
Stratégie	$Acc(T)$	t_c	$Acc(T)$	t_c
Bandit	$0,822$	1127	$0,995$	637
Semi-Bandit	$0,839$	1957	$0,994$	1157
P-BSB-RE	$0,826$	1692	$0,986$	831
P-BSB-OE	$0,824$	655	$0,983$	517
P-BSB-RD	$0,824$	1420	$0,986$	759

Tableau 4 – Résultats observés en moyenne pour les approches considérées sur **Jester** et **MoviesLens**.

4.4.2 Tests statistiques

Nous réalisons en premier lieu des tests de *Kruskal-Wallis (KW)* afin de metre en évidence les inégalités entre les résultats obtenus par chacun des algorithmes, c.-à-.d., nous testons l'hypothèse nulle H_0 : « Il n'y a pas de différence significative entre les résultats des différentes approches (médianes) ». Si le test de *KW* indique qu'il existe des différences entre les résultats, il sera alors nécessaire de réaliser des tests de *Rang signés de Wilcoxon (RW)* deux à deux sur la précision globale et l'itération de convergence, c.-à-.d., nous testons l'hypothèse nulle H_0 : « Il n'y a pas de différence significative entre les résultats entre chaque paire d'approches appliquées à chaque algorithme ». Par la suite, nous indiquerons donc : si l'hypothèse nulle est rejetée ou non, et la valeur de *p* correspondante pour chaque comparaison que nous observerons.

4.4.3 Analyse des résultats

Itération de convergence : Même si nous observons un léger avantage à employer l'approche **P-BSB-OE**, cela reste en revanche statistiquement non significatif (Tests *KW* : $p > 0.05$) pour les cas contextuels comme non contextuels.

Précision globale : Les tests de *KW* nous indiquent qu'il existe une différence significative entre les mesures de précision globale obtenues par l'application de chacune des 5 approches considérées au travers d'un même algorithme, et cela pour chacun des algorithmes appliqués dans les cas contextuels comme non contextuels ($p < 0.01$). Dans le cas non-contextuel, les trois déclinaisons de **P-BSB** obtiennent une précision globale significativement supérieure à l'approche **bandit** (Tests *RW* : $p < 0.01$) et l'approche **Semi-Bandit** obtient une précision globale significativement supérieure aux autres approches (Tests *RW* : $p < 0.01$). Les approches **P-BSB-RD** et **P-BSB-OE** obtiennent des résultats équivalents et les tests de *RW* indiquent qu'ils ne présentent pas de différences significatives (Tests *RW* : $p > 0.05$). Dans le cas contextuel, l'approche **bandit** obtient une précision globale significativement supérieure aux autres approches (Tests *RW* : $p < 0.01$). L'approche **P-BSB-RE** obtient une précision globale non significativement supérieure à l'approche **P-BSB-RD** (Tests *RW* : $p > 0.05$). L'approche **P-BSB-RE** quant à elle obtient des résultats significativement supérieurs à l'approche **P-BSB-OE** (Tests *RW* : $p < 0.01$).

Observations : Ces résultats sont obtenus alors que dans les meilleurs cas, c-à-d lors des itérations ou ρ prend sa valeur maximale : 4, **P-BSB** n'emploie que 40% des retours utilisateurs considérés par les stratégies **bandit** et **semi-bandit** et que dans les pires cas, c-à-d lors des itérations ou ρ prend sa valeur minimale : 0, aucun retour utilisateur n'est exploitable pour l'apprentissage de l'agent. Nos résultats sur les jeux de données **RS-ASM**, **Poker Hand** et **Covertype** avec les algorithmes *COM-MAB* confirment les tendances observées lorsque la part de récompenses non observée est moins importante ($k = 3$ et $0 \leq \rho \leq 3$) et nous permettent de confirmer l'adéquation de notre approche dans ce type d'application.

Conclusion : À la vue des résultats expérimentaux, l'objectif applicatif visé par notre approche - acquérir une précision globale proche de celles obtenues avec les approches classiques malgré un nombre de retours utilisateur restreint, voire inexistant à certaines itérations - est atteint par les variantes de **P-BSB** proposées.

5 Conclusions et Perspectives

Notre objectif final est d'intégrer un système de recommandations guidé par les besoins utilisateurs en environnement marin où un vecteur complet de récompenses R_t serait difficile à observer.

Ainsi, dans cet article, nous avons proposé et appliqué une approche combinatoire à plusieurs algorithmes de bandits-

manchots à tirages simples issus de la littérature. Nous les avons évalués en termes de précision globale et d'itération de convergence sur plusieurs jeux de données du monde réel. Les résultats que nous avons obtenus sont en faveur d'une utilisation de l'approche combinatoire pour les systèmes de recommandation à choix multiples.
La principale contribution de cet article porte sur la mise au point et l'expérimentation d'une nouvelle méthode de prise en compte du retour utilisateur : **P-BSB**. Cette approche propose trois variantes : 1) **RE** qui observe les récompenses associées aux ρ bras de S_t de plus haute espérance de récompense ; 2) **OE** qui consulte les récompenses des bras de S_t ayant été le moins observés à l'itération t ; 3) **RD** qui emploie une sélection aléatoire de ρ bras parmi S_t. Dans les cadres contextuels comme non contextuels, l'approche partielle combinant les stratégies bandit et semi-bandit offre des performances proches des approches classiques, malgré un nombre restreint de retours utilisateur.
L'acquisition et la valorisation du retour utilisateur constitue un défi majeur dans le domaine de l'apprentissage automatisé et les résultats obtenus par *P-BSB* encouragent des perspectives d'une mise en application réelle pour un apprentissage en ligne. À ce titre, l'une des perspectives imminentes que nous envisageons est d'étudier une approche complémentaire où la stratégie de prise en compte du retour utilisateur serait déterminée dynamiquement par l'agent à chaque itération.

Remerciements

Ces travaux ont été menés par l'entreprise KARA TECHNOLOGY en collaboration avec les laboratoires du LERIA et ESEO-TECH et avec le soutien de l'Association Nationale de la Recherche et de la Technologie (ANRT).

Références

[1] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1, 2012.

[2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pages 127–135, 2013.

[3] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, pages 397–422, 2002.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3) :235–256, 2002.

[5] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SICOMP*, 32(1) :48–77, 2002.

[6] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *ARXIV*, abs/1904.10040, 2019.

[7] Fondation Bénéteau. Les attentes des futurs plaisanciers. *Rapport FIN*, 2014.

[8] Lixing Chen, Jie Xu, and Zhuo Lu. Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. In *NIPS*, pages 3247–3256. Curran Associates, Inc., 2018.

[9] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit : General framework and applications. In *ICML*, volume 28 of *Proceedings of Machine Learning Research*, pages 151–159, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[10] Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *NIPS*, pages 2116–2124. Curran Associates, Inc., 2015.

[11] James A. Grant, David S. Leslie, Kevin Glazebrook, and Roberto Szechtman. Combinatorial multi-armed bandits with filtered feedback. *ARXIV*, 2017.

[12] Nicolas Gutowski. *Context-aware recommendation systems for cultural events recommendation in Smart Cities*. Theses, Université d'Angers, November 2019.

[13] Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Improved regret bounds for bandit combinatorial optimization. In *NIPS*, pages 12027–12036. Curran Associates, Inc., 2019.

[14] Luc Jaulin, Fabrice Bars, Benoit Clement, Yvon Gallou, Olivier Menage, Olivier Reynet, Jan Sliwka, and Benoit Zerr. Suivi de route pour un robot voilier. *CIFA*, 07 2012.

[15] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670. ACM, 2010.

[16] Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. Contextual combinatorial cascading bandits. In *ICML*, volume 48 of *Proceedings of Machine Learning Research*, pages 1245–1253, New York, New York, USA, 20–22 Jun 2016. PMLR.

[17] Jonathan Louedec. *Bandit strategies for recommender systems*. Theses, Université Paul Sabatier - Toulouse III, November 2016.

[18] Alexander Luedtke, Emilie Kaufmann, and Antoine Chambaz. Asymptotically optimal algorithms for multiple play bandits with partial feedback. *ARXIV*, 06 2016.

[19] H. Robbins. Some aspects of the sequential design of experiments. *Bull. of the AMS*, pages 527–535, 1952.

[20] Karthik Abinav Sankararaman. Semi-bandit feedback : A survey of results. , 2016.

[21] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning : An introduction*, volume 1. MIT press Cambridge, 1998.

[22] M. van Aartrijk, C. Tagliola, and P. Adriaans. Ai on the ocean : the robosail project. *ECAI*, pages 653–657, 2002.