Gestion des logs dans les problèmes de Bandits Contextuels

Emmanuelle Claeys, Myriam Maumy-Bertrand, Pierre Gancarski

Gestion des logs dans les problèmes de Bandits Contextuels

Emmanuelle Claeys [1]& Myriam Maumy-Bertrand[2] & Pierre Gançarski [3]

[1] *IRMA, 7 rue René Descartes, 67084 Strasbourg, claeys@math.unistra.fr*
[2] *IRMA, 7 rue René Descartes, 67084 Strasbourg, mmaumy@math.unistra.fr*
[3] *ICUBE, 300 Bd Sébastien Brant, 67400 Illkirch-Graffenstaden fbertran@math.unistra.fr*

Résumé. Récemment, de nouvelles méthodes prometteuses optimisent les tests A/B en utilisant l'allocation dynamique. Elles permettent d'obtenir un résultat plus rapide pour déterminer quelle variation est la meilleure, ce qui permet à l'utilisateur d'économiser des coûts. Cependant, l'allocation dynamique par les méthodes traditionnelles reste contraignante sur le type de données utilisées. Dans cet article, nous présentons une nouvelle méthode qui permet d'intégrer des séries temporelles (*i.e log*) pour améliorer l'allocation dynamique dans le cadre d'un A/B test. Cet article fournit des résultats numériques sur des données d'essai réelles, pour démontrer l'amélioration apportée par la méthode par rapport aux méthodes traditionnelles.

Mots-clés. Allocation dynamique, bandit multi-bras, A/B test, séries temporelles.

Abstract. Recently, promising new methods are optimizing A/B tests with dynamic allocation. They allow for a quicker result to determinate which variation is the best, saving costs for the user. However, dynamic allocation by traditional methods has constraints on the data used. In this article, we present a new method that allows us to find the best variation in a short period of time using time series. This paper provides numerical results on real test data to demonstrate the improvement of the method over traditional methods.

Keywords. Dynamic allocation, multi-armed bandit, A/B test, Time series.

1 Introduction

Dans le contexte du web marketing, la méthode traditionnelle d'A/B testing consiste à comparer, après une période fixée, deux versions d'une page web ou d'une application afin de déterminer laquelle est la plus performante. La variation originale, généralement dénommée A, est comparée avec des variations alternatives :B, C, ... [1]. Dans une première approche, dite fréquentiste, ces variations sont présentées uniformément aux *visiteurs* du site. Une partie d'entre elles sera alors dirigée vers la première variation tandis que l'autre

1. pour faciliter la lecture, nous nous placerons dans cet article dans le cas d'un test ou deux variations sont possibles : A et B

sera affectée à la seconde. Un test statistique permet par la suite de tester l'efficacité de la version A et B sur différents indicateurs comme le taux de conversion [2]. En d'autres termes, l'*utilisateur* du test va vérifier quelle variation déclenche le plus de clics, d'abonnements, d'achats ...Les résultats permettent alors de déterminer la meilleure stratégie marketing à adopter.

L'allocation du trafic classique demande à l'utilisateur de fixer en amont du test le ratio du trafic associé à chaque variation et une période de test (période d'*exploration*). Par exemple, en choisissant une répartition 50/50 et après quelques semaines, l'utilisateur analyse les résultats. Si les résultats montrent qu'une variation performe mieux qu'une autre (par rapport à une métrique fixée, comme le taux de conversion), l'utilisateur change manuellement la répartition du trafic (période d'*exploitation*).

Une autre approche est possible, reposant sur l'allocation dynamique. Utilisant un algorithme de bandit manchot (décrit section 2), l'allocation dynamique du trafic va, au cours du test, assigner aux nouveaux visiteurs, la variation la plus performante. L'allocation dynamique permet de s'affranchir d'un temps d'exploration arbitrairement choisi par l'utilisateur. En explorant le plus rapidement possible, tout en étant certain de trouver la meilleure variation, l'utilisateur minimise ses pertes (engendrées par une variation sous optimale). Lorsque la récompense dépend du contexte, c'est à dire des caractéristiques du visiteur testé (par exemple son âge, son sexe *etc.*). Les méthodes d'allocation dynamique adoptent une stratégie d'allocation en fonction de ses caractéristiques. Malheureusement, les méthodes actuelles sont restrictives sur le type de contexte : il ne peut être défini que par un vecteur comportant des valeurs numériques. Ces restrictions limitent l'utilisation des données dont dispose l'utilisateur. En effet ce dernier peut avoir un historique pour chaque visiteur, à sa disposition (c'est à dire des *logs*). Cet article propose une nouvelle méthodologie et un algorithme associé pour permettre une allocation dynamique contextuelle lorsque le visiteur est défini par plusieurs séries temporelles binaire où numérique. La section 2 présente le formalisme associé aux modèles d'allocations dynamiques et les limites de ces modèles et les outils permettant d'utiliser les séries temporelles comme des contextes et la section 3 présente notre contribution et les résultats suite aux expériences réalisées. La section 4 conclue sur ces résultats et propose des perspectives possibles.

2 Le Problème des Bandits Manchots

Le problème des bandits manchots offre un cadre théorique à la prise de décision itérative dans un environnement incertain et réalise une exploration adaptée à la différence entre deux variations. Les algorithmes de bandits offrent de bonnes performances (Lattimore & Szepesvári, 2019) lorsqu'un utilisateur est face au dilemme d'*exploration* (estimer la récompense potentielle de chaque variation) et d'*exploitation* (proposer la

2. le taux de conversion mesure ici le rapport entre les visiteurs ayant réalisé l'action recherchée dans le cadre du test et le nombre total de visiteurs soumis au test.

variation la plus performante de manière à maximiser ses gains). Introduit par (Lai & Robbins, 1985), un agent dispose d'un ensemble $\mathcal{A} = \{a_1, ..., a_K\}$ contenant $K \in \mathbb{N}$ variations (définies comme des *bras*, par analogie aux machines à sous). A chaque itération $t \in \{1, .., T\}$ avec $T \in \mathbb{N}$ l'agent choisit un bras a_t et obtient une récompense $X_{a_t} \in \mathbb{R}$. Dans le problème de bandits dit "stochastique" comme pour ceux traités par l'algorithme UCB, (Auer, Cesa-Bianchi, & Fischer, 2002) les récompenses X_{a_t} sont tirées depuis une distribution stationnaire de moyenne $\mu_{a_t} \in \mathbb{R}$.

Si nous considérons le problème de bandits manchots comme un problème d'estimation, trouver le bras optimal revient à trouver celui qui maximise son espérance totale de gain ; ce qui peut être insuffisant dans le cas où cette espérance peut être conditionnelle. C'est lorsque cette espérance dépend de covariables (définie par la suite) que le problème de bandits contextuels fut introduit par (Langford & Zhang, 2007).

Dans le cas du bandit contextuel, la récompense obtenue par chaque bras dépend ici d'un contexte $c_t \in \mathbb{R}^d$ (un ensemble de d covariables), présenté à l'agent avant qu'il ne choisisse un bras. Sous hypothèse de stationnarité, la récompense est tirée depuis une distribution de moyenne $E[X_{a_t}|c_t, a_t]$. Une approche pour résoudre le problème est de construire un modèle permettant de prédire la dépendance entre bras, contextes et récompenses. Plusieurs algorithmes utilisés en apprentissage supervisés ont déjà été adaptés au problème de bandits contextuels, notamment les modèles linéaires (Kakade, Shalev-Shwartz, & Tewari, 2008 ; Li, Chu, Langford, & Schapire, 2010). C'est généralement l'algorithme LIN-UCB (Langford & Zhang, 2007), qui est le plus utilisé, en raison de sa faible complexité. Ce modèle de bandit repose sur une régression linéaire empirique, utilisant les contextes et récompenses passée de chaque bras.

2.1 La création des contextes

Les algorithmes présentés dans la section précédente possèdent des restrictions concernant le format du contexte, celui-ci devant être présenté sous forme de vecteur de réels à l'algorithme. Malheureusement, ces restrictions limitent l'utilisation des algorithmes de bandits pour plusieurs utilisations pratiques.

Par exemple, dans le contexte de l'A/B testing dédié au web, supposons que chaque visiteur du site, arrive avec un contexte constitué d'un ensemble de données chronologiques décrivant l'évolution de différents indicateurs depuis sa première prise sur le site (des *logs*).

Une méthode possible pour utiliser ces logs comme contexte est lister la taille maximale $m_1, ..., m_d$ pour toutes les séries représentant l'évolution des d types de covariable. Il faut ensuite définir vecteur de taille $\sum_{i=1}^{d} m_i$ qui sera défini comme le vecteur contextuel de chaque visiteur. Une telle approche pose plusieurs problèmes, notamment le que modèle du bandit contextuel repose sur une modèlisation statistique empirique prenant en entrée un contexte pour fournir une probabilité de succès ou une récompense moyenne. Un contexte de grande taille introduit une grande variabilité sur les prédictions du modèle. Il devient alors nécessaire de collecter beaucoup de données.

Pour répondre au problème du vecteur de contexte de très grande taille, (Bastani & Bayati, 2015) ont construit empiriquement un modèle statistiques de type LASSO pour l'intégrer dans une problématique de bandit. Cependant, cette méthode nécessite que les covariables soit indépendantes entre elles, ce qui n'est pas le cas lorsqu'un ensemble de covariables représentent une série temporelle. Construire un contexte en ligne est une tâche difficile sans à priori sur les données et il n'existe pas, à notre connaissance, d'algorithmes de bandits contextuels pouvant s'adapter à des contextes de taille variable. Une solution pour réduire la dimension d'une série temporelle est le clustering.

Si l'on souhaite comparer des séries à échantillonnages irréguliers ou bien de tailles différentes, une attention particulière doit être portée sur le choix de la mesure de similarité dans le clustering. Parmi les méthodes existantes dans la littérature, on trouve notamment : Dynamic Time Wrapping (D.T.W) (H. Sakoe, 1971), une mesure où les séries temporelles sont déformées par transformation non-linéaire de la variable temporelle, pour observer des états identiques entre deux séries, malgré un décalage dans le temps. Dans cet article, nous nous plaçons dans le cas où potentiellement l'observation des variables de contexte est irrégulière. En choisissant la métrique D.T.W, toutes les données sont utilisées, nous gérons des comportements décalés et un échantillonnage irrégulier en comparant des séquences de longueurs différentes. Une méthode de clustering de type k-means, basée sur le barycentre (D.B.A) et introduite par (Petitjean, Masseglia, Gançarski, & Forestier, 2011) a montré de bonnes performances avec la distance D.T.W.

3 Contribution

Dans cet article, nous souhaitons répondre au désir de l'utilisateur qui souhaite intégrer les séries temporelles dans son algorithme d'allocation dynamique via l'algorithme DBA-LinUcb. Cet algorithme est une version "enrichie" de l'algorithme Lin-Ucb. Dans une étape off-line, on dispose d'une variation originale (page web A) et des logs de visiteurs, collectés avant qu'ils n'arrivent sur la page test. Ces données constituent les données d'apprentissage nécessaires à l'algorithme DBA-LinUcb. Toujours dans cette étape offline, DBA-LinUcb détermine des clusters par une approche non supervisée. Ces clusters sont définis par leurs centroïdes et utilisent la méthode D.B.A. Ces centroïdes seront utilisés par la suite dans l'étape online. L'étape online correspond au démarrage du test A/B. Lorsque chaque visiteurs arrive, l'algorithme DBA-LinUcb l'associe au plus proche centroid d'un cluster (selon la distance DTW). Son cluster devient alors une covariable et un modèle de bandit contextuel : Lin-Ucb utilise cette information pour choisir la variation à affecter. Le schéma général 1 résume notre approche.

3.1 Expérimentation

Les données ont été collectées suite à un A/B test fréquentiste réalisé par un site marchand sur une durée de 35 jours en 2018. Les données incluent des logs associés à

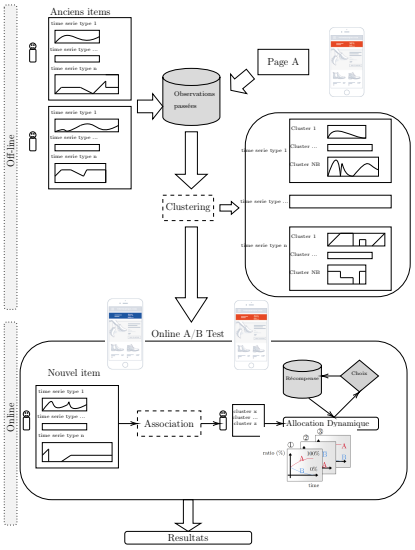

FIGURE 1 – Approche offline/online pour intégrer des séries temporelles aux A/B tests

$T = 11389$ visiteurs différents qui ont navigué sur le site web et ont vu la page test concernée au moins une fois. Chacun d'entre eux est identifié par un identifiant unique : la première fois qu'il arrive sur le site un identifiant lui est généré.

Lorsqu'il arrive sur la page test, une variation (A ou B) lui est attribuée (aléatoirement). Chaque fois qu'un visiteur revient sur la page testée, la même variation lui est affichée. Si un visiteur achète après avoir vu la page test, la récompense est de 1, et 0 sinon. De sa première visite jusqu'à son arrivée sur la page test, sa navigation est enregistrée. On associe à chaque visiteur trois séries temporelles, de même taille (égale au nombre de jours entre sa première et dernière visite avant de voir la page test) :

- `presence_time_serie` : constituée de valeurs binaires. Pour chaque jour où le visiteur s'est rendu sur le site, une valeur 1 est définie, 0 sinon. Chaque série commence et finit par la valeur : 1.
- `hour_time_serie` : constituée de valeurs entières comprises dans $\{0, 1, 2, .., 23, 24\}$. Pour chaque jour où le visiteur s'est rendu sur le site, une valeur égale à son heure d'arrivée sur le site est définie, 0 s'il ne s'y est pas rendu.
- `session_time_serie` : constituée de valeurs comprises dans $\mathbb{R}$. Pour chaque jour où le visiteur s'est rendu sur le site, une valeur égale à sa durée d'activité (en microseconde) sur le site est définie, 0 s'il ne s'y est pas rendu.

On considère dans cette expérience que les logs de visiteurs utilisés dans la phase offline et online sont identiques : nous ne traiterons donc pas de la qualité des clusters. Nous définissons les nombres $k \in \mathbb{N}$ de clusters possibles pour chaque série (choisis via l'indice (L. Davies & Bouldin, 1979)) : (i) $k_{\text{presence\_time\_serie}} = 5$, (ii) $k_{\text{hour\_time\_serie}} = 10$, (iii) $k_{\text{session\_time\_serie}} = 5$.

Pour vérifier si l'intégration des séries temporelles améliore notre modèle de bandits,

nous comparons notre modèle DBA-LinUcb avec un modèle Lin-Ucb, n'utilisant pas de série temporelle elle même, mais uniquement la moyenne des valeurs d'une série. Les visiteurs de la base de données utilisées ayant été assignées à l'une ou l'autre variation possible, on procède par *rejection sampling* : lorsque l'algorithme choisi une variation qui n'a pas été réellement assignée à un visiteur, l'algorithme ignore ce visiteur pour passer au suivant. Après 100 expérimentations sur le dataset mélangées aléatoirement, la performance des modèles se fait par rapport au taux de clic moyenne et est rapporté dans le tableau 1. Nous multiplions le dataset par 2 et par 5 pour observer comment l'écart entre les algorithmes se creuse. Les meilleurs résultats sont surlignés.

Table 1 – Résultats de DBA-LinUcb et Lin-Ucb (avec écart type σ après 100 expérimentations

Algo	Taille	Moyenne $\frac{\text{clic}}{\text{visiteur}}$
DBA-LinUcb	11389	**0.126** σ : 0.003
Lin-Ucb	11389	0.124 σ : 0.0023
DBA-LinUcb	11389*2	**0.126** σ : 0.002
Lin-Ucb	11389*2	0.124 σ : 0.0012
DBA-LinUcb	11389*5	**0.130** σ : 0.002
Lin-Ucb	11389*5	0.124 σ : 0.0011

4 Conclusion

Dans le contexte d'A/B testing sur du webmarketing, il est difficile d'observer des grandes différences sur la métrique d'intérêt (ici le taux de transactions). En effet, sur un site web, pour beaucoup de visites, il y a généralement peu d'acheteurs. Cependant, l'intégration des séries temporelles dans une approche d'allocation dynamique permet à la fois de trouver quelle est la meilleure variation pour un contexte donné (l'utilisateur regardera les coefficients associés à chaque cluster, et comparera ces coefficients pour les deux variations), et de maximiser ses gains . Si l'utilisateur avait proposé la variation ayant le plus grand taux de transaction, indépendamment du contexte, l'utilisateur n'aurait pas obtenu une taux supérieur à 0.126, quelque soit la taille du dataset. L'algorithme DBA-LinUcb indique de meilleurs résultats lorsque l'expérience se prolonge dans le temps (dataset multiplié). Des expériences supplémentaires sont actuellement en cours et paraissent confirmer cette hypothèse. De plus nous testons d'autres algorithmes que Lin-Ucb et les résultats apparaissent comme très encourageants.

Références

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed

bandit problem. *Machine Learning*, *47*(2-3), 235-256.

Bastani, H., & Bayati, M. (2015, 01). Online decision-making with high-dimensional covariates. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2661896

H. Sakoe, S. C. (1971). A dynamic programming approach to continuous speech recognition. *Proceedings of the Seventh International Congress on Acoustics*, *3*, 65–69.

Kakade, S. M., Shalev-Shwartz, S., & Tewari, A. (2008). Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on machine learning* (pp. 440–447). New York, NY, USA : ACM. doi: 10.1145/1390156.1390212

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*(1), 4–22. Consulté sur `http://www.cs.utexas.edu/~shivaram`

Langford, J., & Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. In *Nips.*

Lattimore, T., & Szepesvári, C. (2019). *Bandit algorithms*. Cambridge University Press (preprint).

L. Davies, D., & Bouldin, D. (1979, 05). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *PAMI-1*, 224 - 227. doi: 10.1109/TPAMI.1979.4766909

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. *CoRR*.

Petitjean, F., Masseglia, F., Gançarski, P., & Forestier, G. (2011). Discovering significant evolution patterns from satellite image time series. *International Journal of Neural Systems*, *21*(06), 475-489. Consulté sur `http://www.worldscientific.com/doi/abs/10.1142/S0129065711003024` (PMID : 22131300) doi: 10.1142/S0129065711003024