Apprentissage machine orienté QoS pour l'accès opportuniste au spectre

Navikkumar MODI[1], Philippe MARY[2], Christophe MOY[1]

[1] CentraleSupelec, IETR, UMR CNRS 6164
Avenue de la Boulaie, Cesson-Sévigne, France 35576

[2]INSA, IETR, UMR CNRS 6164
20 Avenue des Buttes de Cöesmes, F-35043 Rennes, France
navikkumar.modi@centralesupelec.fr, philippe.mary@insa-rennes.fr
christophe.moy@centralesupelec.fr

Résumé – Dans ce travail, nous nous intéressons au problème de l'apprentissage machine pour l'accès opportuniste au spectre (AOS). Nous présentons un nouvel algorithme d'apprentissage machine basé sur la classe des UCB (Upper Confidence Bound). Cet algorithme permet non seulement d'apprendre les canaux qui sont le plus souvent libres mais également ceux qui possèdent la meilleur qualité en terme de rapport signal à bruit (RSB), ce qui permet d'introduire une notion de qualité de service (QoS). La stratégie proposée permet à un utilisateur secondaire de pondérer de manière appropriée un critère particulier associé au canal qu'il choisit, comme la puissance d'émission ou le débit par exemple, en sélectionnant deux coefficients d'explorations différents. Notre algorithme permet un compromis entre l'exploration et l'exploitation lorsque le scénario AOS est modélisé comme un problème Markovien de bandit manchot (BM) stable. Nous montrons que notre algorithme surpasse le traditionnel UCB1 en termes de regret et de choix du meilleur canal.

Abstract – This paper deals with the problem of machine learning for the opportunistic spectrum access (OSA) problem. A new index-based machine learning algorithm extended from the upper confidence bound (UCB) class algorithms is proposed. The newly proposed policy selects a frequency band for transmission which is optimal not only in terms of vacancy but also in terms of quality as well, i.e. signal to noise ratio (SNR). The proposed policy allows secondary users to give appropriate weight to their desired criterion, such as emission power or data rate, by selecting two distinguishable exploration coefficients. Proposed policy achieves optimal tradeoff between exploration and exploitation when the OSA scenario is modeled as a rested Markov multi-armed bandit (MAB) problem. In cognitive radio context, the proposed policy is compared to an existing UCB1 and we show that it outperforms traditional UCB1 in terms of regret and selection of the best channel.

1 Introduction

L'accès opportuniste au spectre (AOS) est une réponse efficace au problème de la pénurie du spectre radio. Dans un contexte AOS, deux types d'utilisateurs sont considérés ; le premier ensemble, que l'on nomme utilisateurs primaires (UP), utilise une technologie radio "licenciée" c'est-à-dire pour laquelle il achète le droit d'utiliser cette partie du spectre. Les utilisateurs secondaires (US) constituent le second ensemble et accèdent au spectre des UPs lorsqu'ils ne l'utilisent pas.

La description du problème AOS par un modèle de bandit-machot (BM) a récemment été utilisée avec succès [1, 2]. Les travaux en [3, 4] utilisent en particulier un modèle BM markovien (BMM). Dans la formulation BMM classique, un agent joue sur K machines (les canaux fréquentiels dans un contexte AOS) en une séquence d'itérations en essayant de maximiser son gain à long terme. La récompense produite par chaque bande est modélisée par un processus markovien dont la statistique est *a priori* inconnue du joueur [5, 6]. La plupart des approches traitent de la détection des bandes libres du spectre où le résultat de l'opération de détection d'une bande a deux issues possibles : libre ou occupée. Cependant, les bandes libres ne sont pas toutes équivalentes en terme de "qualité de canal" selon les conditions de propagation dues aux multi-trajets par exemple. Deux transmissions sur deux bandes libres peuvent donc conduire à un débit effectif différent selon les bandes considérées, comme illustré sur la figure 1. D'un point de vue AOS, il serait donc souhaitable de chercher les bandes libres dans lesquelles la qualité de la transmission est la meilleure.

L'objectif est de gérer le dilemme exploration-exploitation défini comme celui de la part accordée à la recherche de la meilleure bande (exploration) et celle accordée à son utilisation (exploitation). L'idée est donc de concevoir une méthodologie d'apprentissage qui recherche des opportunités de transmission dans un ensemble de bandes libres permettant la meilleur QoS possible. Les auteurs en [7, 8, 9, 10] se sont intéressés à l'introduction d'une information de qualité dans les schémas d'apprentissage machine par renforcement, mais souffrent d'un temps de convergence important. Ce papier présente un nouvel algorithme conçu spécifiquement pour chercher les canaux

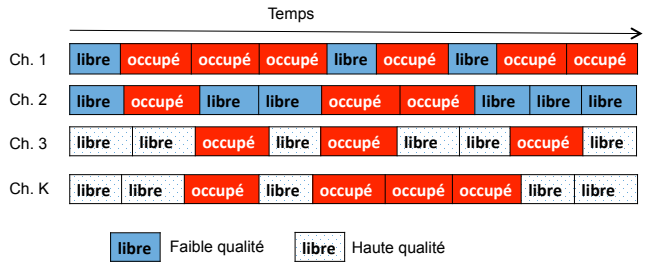

FIGURE 1 – Etat et condition du canal pour l'US.

libres et présentant la meilleur qualité que l'on suppose reliée au RSB.

Dans la partie 2 le modèle BMM stable est présenté, la partie 3 présente une nouvelle procédure d'apprentissage machine basée sur la détection de bande libre ainsi que l'information de qualité du canal. Les résultats de simulations sont présentés dans la partie 4 et nous concluons en 5.

2 Modèle et formulation du problème

2.1 Moteur cognitif

Un équipement RC contient trois fonctionalitées essentielles : un agent de détection de spectre, un agent de prise de décision et un moteur d'apprentissage [1]. Le détecteur de spectre peut être un détecteur d'énergie ou une opération plus sophistiquée comme la détection de la cyclostationnarité. l'US teste d'abord le spectre disponible et décide de transmettre ou non en fonction de la sortie du détecteur d'énergie. Le but du moteur d'apprentissage est alors de prédire le prochain canal à écouter à l'instant suivant. La sélection de la prochaine bande se fait en prenant en compte son état enregistré jusqu'à présent (libre ou occupé) ainsi que le RSB expérimenté sur le canal.

2.2 Modèle mathématique

On considère un ensemble $\mathbb{K} = \{1, \cdots, K\}$, de K canaux. L'évolution des états des bandes est modélisé par un BMM stable, ce qui signifie que l'action de l'US sur le spectre des UP ne modifie pas l'état des canaux autres que celui qui est en train d'être joué par l'US. En effet, l'US n'a aucune raison de modifier l'activité du primaire et de plus, étant seul, la modification de l'état du canal k (i.e. libre à occupé) ne concerne pas d'autres US. La prise en compte de plusieurs US impliquerait la modélisation du problème par un BMM dynamique où l'action des US modifie les états des canaux.

La récompense associée à l'état $q \in S^i$ d'une bande i au temps t est notée $r_q^i(t) \in \mathbb{R}$. $S^i(t)$ est la valeur de l'état de la bande i observée au temps t, i.e. libre ou occupé. Une bande est modélisée par une chaîne de Makov apériodique et irréductible à deux états de matrice de transition $P^i = \left\{p_{k,l}^i, k, l \in S^i\right\}$. La distribution stationnaire de la chaîne de Markov est telle que $\pi_q^i(t) = \pi_q^i \ \forall t$.

La sélection de la bande i à chaque instant t se fait en fonction des états précédemment observés et de la qualité instantanée du canal, $R_q^i(t)$, qui est une fonction du RSB instantanée entre le transmetteur et récepteur secondaire. Le RSB instantané pour un état q est un processus aléatoire stationnaire au sens large à l'ordre 2. La récompense moyenne μ^i, sur une bande i, sous hypothèse de distribution stationnaire π_q^i est :

$$\mu^i = \sum_{q \in S^i} q R_q^i \pi_q^i \tag{1}$$

Dans le cadre d'un BMM, la performance d'une stratégie d'apprentissage est mesurée avec la notion de regret $\Phi(n)$, défini comme la différence entre la stratégie idéale qui jouerait systématiquement le canal avec la plus grande récompense jusqu'à l'instant n et la récompense moyenne de la stratégie mise en oeuvre [6] :

$$\Phi(n) = n\mu^* - \mathbb{E}\left[\sum_{q \in S^i} \sum_{t=1}^{n} R_q^i(t) r_q^i(t)\right] \tag{2}$$

La bande optimale est celle avec la plus grande récompense moyenne μ^* associée à la qualité R_q^*.

3 Algorithme proposé

La stratégie proposée apprend à partir des observations acquises en écoutant les canaux sans en avoir de connaissance statistique. Notre contribution est résumée dans l'algorithme présenté dans la figure 2 ainsi que dans la proposition 1.

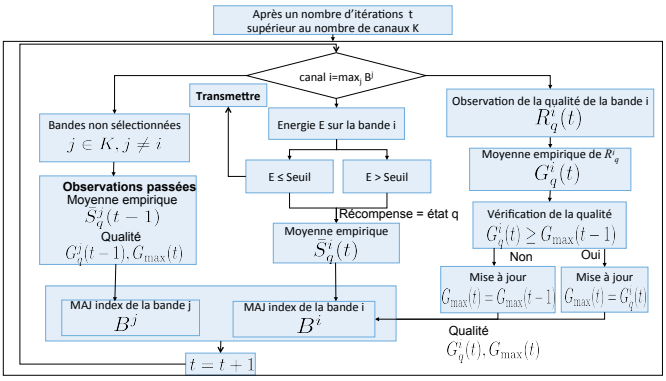

FIGURE 2 – Nouvel algorithme d'apprentissage machine.

L'algorithme commence par tester tous les canaux au moins une fois et l'index de chaque canal, $B^i\left(n, T^i(n)\right)$ défini plus loin, est mis à jour une première fois pour tous les canaux. Une fois le premier parcours du spectre effectué, l'algorithme fait une détection d'énergie sur le canal i qui possède l'index B^i le plus important (branche du milieu de l'algorithme). Si celui-ci est libre, on utilise le canal pour transmettre. Sinon, il n'y a pas d'opportunité de transmission pour l'itération en cours. Une récompense arbitraire $r_q^i(t)$ sur l'état du canal est associée au résultat du détecteur d'énergie et la valeur moyenne de

l'état $\overline{S}^i(t)$ est calculée. En parallèle, la qualité de la bande i est évaluée et mise à jour $G_q^i(t)$ (branche de droite). De plus, B^j $\forall j$, est mis à jour pour toutes les bandes et la prochaine bande à écouter est celle possédant le plus grand index B (branche de gauche). La proposition suivante formalise la stratégie d'apprentissage avec prise en compte de la qualité du canal.

Proposition 1 (QoS-UCB) *Soit un problème AOS décrit par un modèle BMM stable. La loi d'apprentissage du meilleur canal en terme de disponibilité et de qualité s'écrit*

$$B^i(n, T^i(n)) = \overline{S}^i(T^i(n)) - Q^i(n, T^i(n)) + A^i(n, T^i(n)) \tag{3}$$

où $T^i(n)$ est le nombre de fois où le canal i a été testé jusqu'à l'instant n. $\overline{S}^i\left(T^i(n)\right)$ est le terme d'exploitation, $Q^i(n, T^i(n))$ et $A^i(n, T^i(n))$ sont les contributions à l'exploration des autres canaux. On a

$$\overline{S}^i(T^i(n)) = \frac{S^i(1) + S^i(2) + \cdots + S^i(T^i(n))}{T^i(n)}, \forall i \tag{4}$$

où $S^i(T^i(n))$ est l'état de la bande i à l'essai $T^i(n)$, i.e. q_0 ou q_1.

$$Q^i(n, T^i(n)) = \frac{\beta M^i\left(n, T^i(n)\right) \log n}{T^i(n)} \tag{5}$$

où $M^i\left(n, T^i(n)\right) = G_{q,\max}(n) - G_q^i(T^i(n)), \forall i$ et de plus $G_q^i(T^i(n)) = \frac{1}{T^i(n)} \sum_{k=1}^{T^i(n)} R_q^i(k)$ la moyenne empirique des qualités de chaque canaux R_q^i dans l'état q et $G_{\max}^q(n) = \max_{i \in \mathbb{K}} G_q^i(T^i(n))$ est la qualité maximum espérée sur l'ensemble des bandes considérées. Enfin

$$A^i(n, T^i(n)) = \sqrt{\frac{\alpha \log n}{T^i(n)}} \tag{6}$$

où α et β en (5) *et* (6) *sont des paramètres d'exploration.*

La preuve est omise par manque de place, mais on montre que le regret de cette stratégie croit logarithmiquement, comme son homologue UCB classique. L'index B^i comprend donc trois termes. Le premier terme en (4) est la moyenne empirique des états de Markov observés pour la bande i jusqu'à l'instant n. Le terme $Q^i\left(n, T^i(n)\right)$ représente le terme de qualité de la bande testée, dépendant de la récompense $R_q^i(n)$. Enfin $A^i\left(n, T^i(n)\right)$ est le biais permettant d'explorer d'autres canaux lorsque la bande testée est occupée. Trois types de comportement de l'algorithme sont attendus :

1. La bande sélectionnée est optimale en terme de disponibilité et de qualité alors les deux termes en (4) et (5) assurent l'exploitation de cette bande.
2. La bande sélectionnée est optimale en terme de disponibilité ou de qualité. Le terme respectif, (4) ou (5), encourage l'exploitation de cette bande mais les autres termes, i.e. (4) et (6) si le canal est optimal en terme de qualité par exemple, forcent à explorer d'autres bandes.
3. La bande sélectionnée n'est optimale ni en terme de disponibilité ni en qualité, alors le dernier terme (6) force l'exploration des autres bandes.

4 Résultats de simulation

On considère $K = 10$ et $q_0 = 0$ et $q_1 = 1$. De plus, on considère ici que $r_{q_0}^i(t) = 0$ et $r_{q_1}^i(t) = 1, \forall i \in \mathbb{K}$, mais les valeurs peuvent être choisies dans $\mathbb{R}$. Le tableau 1 contient sur les deux premières lignes les probabilités de transition des états des canaux dues à l'activité du primaire et sont arbitrairement choisies. La troisième ligne indique la qualité des bandes disponibles sur le lien de l'US Tx - Rx sur le canal i et dont la connaissance est supposée connue au transmetteur US. De plus R_q^i est une fonction du RSB sur le lien considéré. Enfin la quatrième ligne donne la récompense moyenne de chaque bande évaluée avec (1). La figure 3 compare le regret de notre algorithme avec celui d'un UCB classique, dénommé UCB1 sur la figure, ne prenant pas en compte la qualité du canal [5]. Le coefficient d'exploration de l'UCB1 est pris à $L = 0.5$ comme proposé par les auteurs en [5, 6] pour atteindre les meilleurs performances de convergence. Nous constatons que notre algorithme permet d'atteindre un regret cumulé beaucoup plus faible que le traditionnel UCB1. Cela est du au fait que notre algorithme joue la bande 3 sur le long terme qui est la bande optimale en terme de récompense moyenne, i.e. μ^3. Le regret de l'UCB1 est plus important car il ne donne pas de poids à la qualité du canal $R_{q_1}^i$.

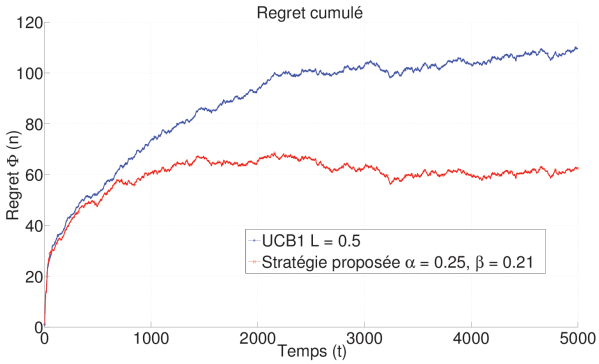

FIGURE 3 – Regret cumulé des deux stratégies.

La figure 4 montre la récompense moyenne cumulée en exécutant la nouvelle stratégie et le classique UCB1. Notre algorithme offre une plus grande récompense comparé à l'UCB1 et donc une qualité de transmission plus grande, en terme de débit par exemple, puisque R_q^i est lié au RSB. Initialement, les deux algorithmes explorent différentes bandes, ainsi la récompense moyenne varie significativement pour les deux. Après un temps de convergence, on voit que la récompense de notre algorithme tend vers 1.45 qui est la qualité de la bande 3 alors que l'UCB1 tends vers 1.22, qualité de la bande 4 qui est la plus souvent libre. La figure 5 confirme cela en montrant l'histogramme des bandes sélectionnées ; notre stratégie privilégie la bande 3 alors que l'UCB1 la bande 4 qui ne possède pas la plus grande qualité et donc qui ne permettra pas une qualité de service optimale sur cette bande.

TABLE 1 – Table des P^i, des qualités $R^i_{q_1}$ et des récompenses moyennes μ^i.

band	1	2	3	4	5	6	7	8	9	10
$p^i_{q_0 q_1}$	0.4	0.5	0.75	0.8	0.6	0.5	0.7	0.1	0.45	0.32
$p^i_{q_1 q_0}$	0.4	0.43	0.25	0.16	0.48	0.77	0.57	0.79	0.67	0.43
$R^i_{q_1}$	1.45	1.77	1.9	1.45	1.30	1.45	1.18	1.25	1.92	1.4
μ^i	0.77	0.99	1.45	1.22	0.76	0.63	0.69	0.22	1.83	0.65

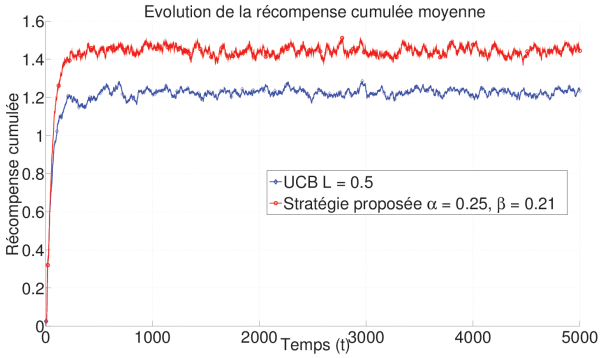

FIGURE 4 – Récompense cumulée des deux stratégies.

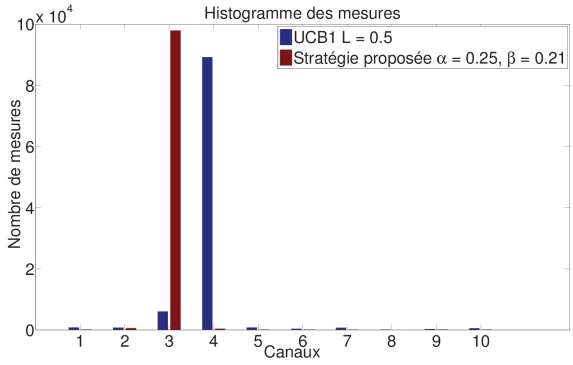

FIGURE 5 – Histogramme comparé de la sélection des canaux

5 Conclusion

Dans cet article, un algorithme d'apprentissage pour l'AOS se basant sur la disponibilité du canal ainsi que sa qualité pour la transmission de l'information a été proposé. L'algorithme a été présenté dans le cadre d'une modélisation du problème AOS par un BMM. La stratégie proposée offre un degré de liberté supplémentaire sur le choix du canal de transmission pour l'utilisateur secondaire. En effet, celui-ci peut décider de pondérer le résultat de la sortie de la détection de spectre par une information sur la qualité du canal pour sa transmission. Les résultats numériques ont montré que la technique proposée est plus performante en terme de regret que l'algorithme UCB standard. De plus, l'apprentissage machine proposée permet de sélectionner le canal avec la meilleur qualité. Nous sommes actuellement en train d'étendre ce travail afin de prendre en compte une métrique d'efficacité énergétique dans le cadre de la RC efficace en énergie.

Remerciements

Ce travail a bénéficié d'une aide de l'Etat attribuée au labex CominLabs et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01.

Références

[1] W. Jouini, D. Ernst, C. Moy *et al.*, "Upper confidence bound based decision making strategies and dynamic spectrum access," *in Proc. ICC'10*, May 2010.

[2] W. Jouini, C. Moy, et J. Palicot, "Decision making for cognitive radio equipment : analysis of the first 10 years of exploration," *EURASIP JWCN,* vol. 2012, no. 26, Jan.

[3] H. Liu, K. Liu, et Q. Zhao, "Learning in a changing world : Restless multi-armed bandit with unknown dynamics," *IEEE Trans. Inf. Theory,* vol. 59, no. 3, 2013.

[4] J. Oksanen, V. Koivunen, et H. V. Poor, "A sensing policy based on confidence bounds and a restless multi-armed bandit model," *CoRR,* vol. abs/1211.4384, 2012.

[5] C. Tekin et M. Liu, "Online algorithms for the multi-armed bandit problem with markovian rewards," *In 48th Annual Allerton Conference,* 2010, pp. 1675-1682.

[6] P. Auer, N. Cesa-Bianchi, et P. Fischer, "Finite-time analysis of the multi-armed bandit problem," *Machine Learning,* vol. 47, no. 2-3, May 2002.

[7] H. N. Pham, J. Xiang, Y. Zhang, *et al.*, "Qos-aware channel selection in cognitive radio networks : A game theoretic approach," *In IEEE Globecom,* Nov 2008.

[8] E. Ahmed, L. J. Yao, M. Shiraz, *et al.*, "Fuzzy-based spectrum handoff and channel selection for cognitive radio networks," *in Proc. IC3INA,* Nov 2013.

[9] A. Ali, M. Iqbal, S. Saifullah, *et al.*, "Qos-based channel and radio assignment algorithm for mesh cognitive radio networks intended for healthcare," *in Proc. MESH*, 2012.

[10] C. Noda, S. Prabh, M. Alves, *et al.*, "Quantifying the channel quality for interference-aware wireless sensor networks," *SIGBED Rev.*, vol. 8, no. 4, dec 2011.