Proposition de stage de M2.
Apprendre en faisant: Approches de type "bandits manchots" pour l'apprentissage et le contrôle de réseaux bayésiens dynamiques.

Encadrement

Nathalie Peyrard et Régis Sabbadin. INRA, Unité de Mathématiques et Informatique Appliquées, Toulouse, équipe Modélisation des Agroécosystèmes et Décision. Email: `nathalie.peyrard@inra.fr, regis.sabbadin@inra.fr`
https://mia.toulouse.inra.fr/Accueil

Mots-clés

Optimisation de la décision ; Apprentissage par renforcement ; Algorithmes de Bandits.

Enjeu finalisé et problématique scientifique

Le changement climatique risque de rendre les systèmes de cultures actuels inadaptés, à la fois économiquement et environnementalement. Il faut imaginer de nouveaux systèmes pour répondre à ce changement et un des leviers disponibles est celui de la rotation des cultures, qui peut être adaptée aux effets du changement climatique sur la production et sur les fonctions fournies par l'écosystème vivant dans les cultures. Mais un nouveau choix de rotation va induire un changement du réseau écologique occupant la culture (certaines espèces peuvent disparaitre, changer de proies ou d'habitat). Ce changement, qui ne sera mesuré précisément que plusieurs années après sa mise en place, aura à son tour un effet sur la production.

Ce problème de conception de rotations innovantes est un problème typique de conception de stratégie de gestion dans un contexte incertain. Il est difficile de trouver la meilleure rotation dans ce contexte car :
1) Etablir la rotation qui maximise (par exemple) la production en se basant uniquement sur la connaissance courante du réseau écologique, peut conduire à une très mauvaise production en pratique.
2) Etablir une rotation qui permette d'apprendre le mieux possible le réseau écologique peut également conduire à une mauvaise production puisque dans ce cas le but n'est plus de la maximiser.

Cette famille de problèmes a généré un fort intérêt pour des approches d'Intelligence Artificielle de type "Adaptive Management", proposant d'intégrer "conduite" et "étude" des écosystèmes, dans le but d'"apprendre en faisant", c'est-à-dire de trouver la décision (rotation ici) qui est le meilleur compromis entre 1) et 2), pour maximiser un objectif (de production, par exemple).

Les méthodes de résolution actuelles ne passent malheureusement pas à l'échelle dans des problèmes où, en même temps, le modèle est incertain et l'espace d'états du système est grand. Une façon de modéliser de manière compacte la dynamique d'un ensemble d'entités en interaction est de se placer dans le cadre des Réseaux Bayésiens Dynamiques (RBD, Dean and Kanazawa, 1989). Avec ce sujet de Master, nous proposons d'explorer de nouvelles méthodes basées sur l'utilisation jointe des RBD pour la modélisation des systèmes et des approches de type *Bandits manchots* (Gittins, 1979) pour l'optimisation de leur gestion, afin de repousser ces limites.

Projet de Master

Durant ce stage, nous considérerons des problèmes de décision où une seule décision de contrôle est prise initialement, puis le processus évolue sur un nombre d'étapes fixé en suivant un modèle de RBD inconnu,

attaché à la décision prise. Plus précisément, nous nous placerons dans le cadre *paramétré* des RBD étiquetés (*RBDE*, Auclair et al., 2017), qui a l'avantage de comporter moins de paramètres à apprendre.

Historiquement, les problèmes d'optimisation de décision dans l'incertain avec modèle mal connu ont d'abord été modélisés dans le cadre des *Bandits Manchots* (Simple Family of Alternative Bandit Processes, SFAB, Gittins, 1979). Cette approche, qui vise à optimiser le choix d'une action dont les effets à venir sont incertains est théoriquement applicable à l'optimisation d'un choix de décision définissant un processus de type RBDE. Une application naïve des méthodes de type SFAB est toutefois inefficace en pratique, compte tenu du nombre de bandits à considérer. Plusieurs extensions ont été proposées (*UCB*, Bubeck and Cesa-Bianchi, 2012 ; *EXP3*, Cesa-Bianchi and Lugosi, 2006), mais elles n'exploitent pas la représentation factorisée spécifique de la loi par un RBDE. L'objectif de ce stage est de tester et d'adapter ces approches dans le cas où la dynamique du système du bandit est supposée suivre un modèle de RBDE inconnu.

Dans un premier temps, l'étudiant retenu s'appropriera le cadre des RBD et RDBE, et effectuera une analyse bibliographique des approches classiques du domaine des bandits manchots. Parmi ces approches, il adaptera celles jugées les plus pertinentes au cas des bandits construits sur des RBDE. Enfin, l'étudiant évaluera expérimentalement ces approches adaptées, sur un problème de choix de rotation des cultures dans une parcelle. En fonction du déroulé du stage, il pourra également être envisagé d'étudier théoriquement les garanties sur les performances de ces méthodes, en s'inspirant des méthodes d'évaluation dédiées aux bandits bayésiens (*Bayes-UCB*, Kaufman et al., 2012 ; *Thomson Sampling*, Agrawal and Goyan, 2012).

Compétences requises

L'étudiant retenu devra présenter de bonnes compétences dans l'un des domaines suivants: (i) Machine Learning, (ii) Apprentissage statistique ou (iii) Apprentissage par Renforcement.

Le sujet étant novateur et riche en développements potentiels, il sera la base d'un sujet de doctorat sur les méthodes d'apprentissage par renforcement pour les problèmes de décision séquentielle dans l'incertain, factorisés. Le candidat devra être motivé par la recherche et posséder un bon dossier universitaire de manière à pouvoir prétendre à une bourse de doctorat.

Durée et rémunération

La durée du stage est de 5 à 6 mois (à définir avec l'étudiant) et la rémunération d'environ 550 Euros par mois.

Bibliographie

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proc. of the 25th Conference On Learning Theory.*

Auclair, E., Peyrard, N., Sabbadin, R. (2017). Labeled DBN learning with community structure knowledge. In *Proc. of the 27th European Conference on Machine Learning.*

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems, *Foundations and trends in machine learning*, 5(1):1-122.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*, Cambridge University Press.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5:142–150.

Gittins, J. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Proc. of the 23rd conference on Algorithmic Learning Theory.*