Un algorithme « bandit manchot » pour le choix de nouvelles situations d'apprentissage à l'intérieur d'un environnement virtuel

Y. Bourrier[1], Vanda Luengo[1] F. Jambon[2], Catherine Garbay[2]

[1]Sorbonne Université, LIP6 – équipe MOCAH
[2] Université Grenoble Alpes, LIG

Yannick.bourrier@imag.fr

Résumé

Nos recherches s'intéressent à formation de techniciens aux compétences non-techniques (CNT) lors de situations critiques, par l'entremise d'un environnement virtuel (EV). Ce résumé présente le modèle utilisé pour générer dynamiquement, et de manière personnalisée, les situations critiques les plus pertinentes pour l'apprentissage de CNT, en fonction des caractéristiques de chaque apprenant. L'approche employé dans cet objectif se base sur un algorithme de type « bandit manchot », qui considère le choix d'une nouvelle situation d'apprentissage comme un problème d'exploration/exploitation. Le modèle alterne les phases d'exploration, pour trouver les situations les plus adaptées pour l'apprenant à un moment donné- et les phases d'exploitation, pour mettre à profit les connaissances découvertes sur les situations afin de faire progresser l'apprenant.

Mots Clef

Environnements Informatiques pour l'apprentissage humain, prise de décision, apprentissage par renforcement.

1 Introduction

Dans la plupart des milieux présentant un degré de risque (e.g. : la conduite, la médecine, l'aviation…), on a constaté que lors de l'apparition de situations critiques, des techniciens expérimentés possédant en théorie toutes les compétences requises pour éviter une catastrophe, se montraient incapables d'effectuer les bons gestes [1]. Une explication à ces erreurs peut se trouver du côté des compétences non-techniques, c'est-à-dire, d'un panel de compétences métacognitives telles que la conscience de la situation, la prise de décision, la communication, le travail en équipe, ou la gestion du stress. Afin d'améliorer la sécurité dans ces domaines, il convient de former les techniciens à l'application de telles compétences. Cette tâche de formation est complexe, car ces CNT s'acquièrent majoritairement de manière empirique, et sont mobilisées avant tout lors de situations critiques [2]. Les environnements Informatiques pour l'Apprentissage Humain (EIAH) sont particulièrement adaptés à la formation aux CNT, car ils permettent de simuler des contextes favorables à leur entrainement, sans pour autant générer des risques humains toujours présents lors de situations critiques réelles.

L'un des buts de notre EIAH est de générer des situations adaptées aux particularités de chacun. Pour ce faire, les actions et perceptions de chaque apprenant sont analysées en temps réel, puis fournies à un modèle de « diagnostic des compétences », qui fournit en retour une photographie des CNT de ce dernier (sous la forme d'un réseau bayésien) [3]. A partir de cette photographie, un modèle décisionnel explore la base des situations d'apprentissage disponibles, et sélectionne la plus adaptée à l'apprenant. L'apprenant fait face à cette nouvelle situation, un nouveau diagnostic est fourni, qui est à son tour utilisé pour la génération d'une nouvelle situation.

L'objectif central est de faire progresser l'apprenant en lui fournissant toujours des situations pertinentes pour lui, c'est-à-dire d'une difficulté adaptée, et ciblant principalement ses CNT les plus faibles (le modèle cherche d'abord à renforcer les faiblesses de chacun). Le choix d'une nouvelle situation est réalisé en combinant les informations obtenues par le diagnostic, à un algorithme de type « bandit manchot ». Dans le problème du bandit manchot, un joueur dans un casino fait face à un certain nombre de machines à sous. Chaque machine à sous (i.e. : chaque bras) donne une récompense moyenne inconnue, et l'objectif du joueur est de trouver le bras lui donnant la récompense la plus forte. Pour ce faire, il lui faut explorer les différents bras pour acquérir des connaissances sur les récompenses que ces derniers fournissent, puis exploiter les connaissances acquises afin de maximiser le gain.

Dans notre cas, et de manière inspirée par [4], notre modèle de diagnostic des CNT joue un double rôle. Tout d'abord, il effectue un premier filtrage sur la totalité des situations disponibles, pour sélectionner seulement celles d'une difficulté proche du niveau estimé de l'apprenant. Ensuite, il joue le rôle du joueur face aux bandits, et les situations préalablement sélectionnées constituent les bras. L'objectif est de trouver, à chaque instant, le bras maximisant le gain d'apprentissage de chaque apprenant, tel qu'observé par l'entremise du modèle de diagnostic. Le modèle va explorer les situations disponibles pour obtenir une estimation des récompenses qu'elles fournissent (i.e. : le gain d'apprentissage), puis exploiter celles qui fournissent le meilleur gain. Cette tâche d'exploration/exploitation comporte deux difficultés intrinsèques au domaine des EIAH :

- Les récompenses fournies par chaque bras changent au fil du temps. En effet, à mesure que les compétences d'un apprenant augmentent, certaines situations auparavant

très adaptées fourniront une récompense moindre, tandis que d'autres situations seront sélectionnées comme pertinentes par le modèle de diagnostic et devront être explorées.

- Le joueur étant un modèle de diagnostic estimant de manière probabiliste les compétences de l'apprenant, et non l'apprenant lui-même, les conséquences du choix d'une situation ou d'une autre sur ce dernier sont soumises à un degré d'incertitude.

Plusieurs algorithmes sont à l'heure actuelle testés, notamment un algorithme de type *epsilon-greedy* qui explore d'abord fortement les situations, puis exploite fortement la meilleure ensuite, et un algorithme probabiliste de type *softmax*, où chaque bras est tiré avec une probabilité correspondant à sa récompense estimée mise au rapport des récompenses estimées de tous les autres bras. Ces deux algorithmes sont confrontés à d'autres approches, comme par exemple, une séquence de scénarios déterminée *a priori* par un expert. Les premiers tests semblent indiquer l'efficacité des algorithmes de type bandit manchot, et notamment de l'approche *softmax*, pour produire des séquences de scénarios personnalisées, et de manière générale plus efficaces pour l'apprentissage qu'une séquence experte.

Bibliographie

[1] Mitchell, M. L., & Flin, R. (Eds.). (2012). Safer surgery: analysing behaviour in the oper- ating theatre. Ashgate Publishing, Ltd

[2] Bourrier, Y., Jambon, F., Garbay, C., & Luengo, V. (2016, September). An Ap-proach to the TEL Teaching of Non-technical Skills from the Perspective of an Ill-Defined Problem. In European Conference on Technology Enhanced Learning (pp. 555-558). Springer International Publishing.

[3] Bourrier, Y., Francis, J., Garbay, C., & Luengo, V. (2018, June). A Hybrid Architecture for Non-Technical Skills Diagnosis. In Intelligent Tutoring Systems.

[4] Clement, B., Roy, D., Oudeyer, P. Y., & Lopes, M. (2013). Multi-armed bandits for intelligent tutoring systems. arXiv preprint arXiv:1310.3174.