Evaluer les apprentissages : comparaison de deux techniques de diagnostic de connaissance

Marie Sacksick

Evaluer les apprentissages : comparaison de deux techniques de diagnostic de connaissance

Marie Sacksick[1]

[1] Université Paris 8, CHArt – France
& Universidad de Cordoba, KDIS – Espagne
& Domoscio – France
marie.sacksick@ntymail.com

Résumé. L'item Response Theory (IRT) et le Bayesian Knowledge Tracing (BKT) sont deux modèles d'évaluation de la maîtrise d'une connaissance par un apprenant. Ces deux modèles sont généralement comparés sur la précision de la prédiction d'une bonne réponse d'un apprenant à une question. Nous montrons dans cet article que ces deux modèles s'adressent à des situations distinctes pour lesquelles chacun d'eux peut être utilisé, et nous émettons l'hypothèse que cette hiérarchisation par la précision évolue en fonction de la quantité de réponses de l'apprenant.

Mots-clés. Théorie de la Réponse à l'Item, Bayesian Knowledge Tracing, Système Tuteur Intelligent, comparaison de modèles, modèle prédictif, modèle explicatif.

Abstract. Item Response Theory (IRT) and Bayesian Knowledge Tracing (BKT) are two evaluation models of the mastery degree of a concept by a learner. In general, these two models are compared based on the prediction prevision of the correct answer of a learner to a question. We show in this article that these models tackle different situations for which each of them can be used, and we hypothesize that this hierarchy by the precision tends to evolve according to the amount of questions the learner answered to.

Keywords. Item Response Theory, Bayesian Knowledge Tracing, Intelligent Tutoring System, model comparison, predictive model, explicative model.

1 Introduction

Lors de la dernière décade, le domaine des systèmes de tuteur informatisé (ITS, Intelligent Tutoring Systems) est entré dans une phase d'opérationnalisation et d'industrialisation, avec des entreprises comme Knewton aux Etats-Unis ou Domoscio en France. Dans un ITS, pour fournir une aide adaptée qui favorise l'apprentissage, il est nécessaire d'estimer la maîtrise d'une connaissance par un apprenant. Le choix de la technique estimant la maîtrise d'une connaissance est sujet à discussion aussi bien dans le milieu de la recherche que dans le milieu industriel.
Cette maîtrise est généralement traduite par la probabilité d'obtenir une réponse correcte à une question, cette probabilité pouvant dépendre de la question elle-même.

Ce besoin d'estimation se retrouve dans d'autres outils en support à l'apprentissage humain comme dans le système RiARiT [1].
Lorsque l'on a besoin d'estimer la maîtrise d'une connaissance par un apprenant, il faut sélectionner parmi les techniques de diagnostic qui permettent une telle estimation. Pour cela, il est courant de choisir une technique pour sa précision, c'est-à-dire sa capacité à prédire correctement si un apprenant va répondre correctement ou non à une question. Nous souhaitons souligner dans cet article d'autres critères et des spécificités aux situations, et comparer les plus connus des modèles que sont l'Item Response Theory (IRT, théorie de la réponse à l'item) et le Bayesian Knowledge Tracing (BKT). Il en existe également d'autres comme le Performance Factor Analysis [2] ou encore le Knowledge Tracing Machine [3].
Dans un premier temps nous présenterons plus avant les modèles IRT et BKT, avant de les mettre en regard dans la troisième partie.

2 Spécificités des modèles IRT et BKT

Dans cette partie, nous présentons les modèles IRT et BKT, en particulier les idées et les hypothèses sur lesquelles ils reposent.

2.1 IRT et modèles associés

L'Item Response Theory (IRT) repose sur un modèle probabiliste qui suppose une relation entre la validité de la réponse de l'apprenant à une question d'un côté et les caractéristiques de la question et la maîtrise du concept par l'apprenant de l'autre.
Etant donnés un apprenant A_j et une question Q_i, la probabilité de succès de l'apprenant à la question s'écrit :

$$P_i\big(obs = correct \big| A_j\big) = g(A_j, Q_i) \quad \textbf{(1)}$$

Dans le modèle 2-PL (2 Parameters Logistic model), la fonction g peut s'écrire :

$$g\big(A_j, Q_i\big) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad \textbf{(2)}$$

Où θ_j est le niveau de maîtrise de la connaissance sur lequel repose la question par l'apprenant ; a_i est le niveau de discrimination de la question ; b_i représente le niveau de difficulté de la question.
Disposant des données de N_j apprenants ayant répondu à N_i questions, il est possible d'estimer la valeur de ces paramètres grâce à la minimisation de la vraisemblance ; une fonction basée sur la probabilité décrite en (1).
Le modèle de l'IRT repose sur plusieurs hypothèses [4], [5], notamment celle de la « stabilité de l'apprenant », qui suppose que le niveau de maîtrise θ est identique à tous les moments où l'apprenant répond aux questions.
La fonction g présentée dans (2) est une des plus connues, mais il existe d'autres possibilités comme le modèle de Rasch qui considère que $a_i = 1$. Il est également

possible d'étendre le modèle afin de considérer que la question dépend de plusieurs concepts ; la probabilité de répondre correctement dépend alors de plusieurs maîtrises [4].

2.2 BKT et modèles associés

Le Bayesian Knowledge Tracing (BKT) est un modèle basé sur les chaînes de Markov cachées dans lequel on trouve deux états : un état de maîtrise de la connaissance, et un état de non maîtrise. Partant du principe qu'il est impossible de savoir dans lequel de ces deux états se trouve l'apprenant, il est en revanche possible d'avoir des observations, que sont les réponses à des questions, qui donnent des indications sur l'état de l'apprenant. Lorsqu'un apprenant réussit une question, il peut être soit dans l'état de maîtrise, soit avoir deviné (*probabilité « guess »*) et être en réalité dans l'état de non maîtrise. Lorsqu'un apprenant répond incorrectement à une question, il peut effectivement ne pas connaître, ou bien s'être trompé pour des motifs comme la fatigue (*probabilité « slip »*). Répondre à une question offre à l'apprenant l'opportunité de passer de l'état « d'ignorance » à « connaissance » (*probabilité de transition*). Il est également possible que l'apprenant maîtrise déjà le concept (*probabilité initiale* avant le début du test et de ses réponses aux questions). On pourra retrouver toutes les équations associées dans [6].
Plusieurs variantes du BKT ont été proposées afin d'assouplir le modèle, notamment dans [7] où les paramètres de « *slip* » et « *guess* » sont contextualisés, ou dans [8] qui cherche à déterminer le degré de difficulté d'une question et à le prendre en compte dans le calcul des états.

3 Comparaison des deux modèles

Nous considérons dans la suite la situation suivante : dans le cadre d'un outil d'apprentissage adaptatif, nous cherchons à déterminer quelle méthode serait la plus adaptée entre l'IRT et le BKT. Cet outil d'apprentissage adaptatif recommande des contenus pédagogiques aux apprenants en fonction de leurs objectifs, de leur niveau de maîtrise des connaissances, et de leur parcours sur la plateforme d'apprentissage, et ce jusqu'à ce que ces objectifs pédagogiques soient atteints.

3.1 Fiabilité des modèles

Généralement, les modèles IRT et BKT sont comparés comme des modèles classiques d'apprentissage automatique ; autrement dit, pour plusieurs ensembles de données une méthode de validation croisée est utilisée (entraînement sur une partie des données pour tester sur l'autre partie des données), et la performance est comparée en termes de RMSE (Root Mean Squared Error) et d'AUC (Area Under the Curve).
Selon les variantes des modèles qui sont choisies, et selon les bases de données sur lesquelles elles sont testées, les résultats sont évidemment différents. Ainsi par exemple dans l'article de Minn et al. [9], l'IRT est bien meilleur en termes d'AUC

mais le BKT est un peu au-dessus en termes de RMSE. Ces performances dépendent des jeux et de données et du contexte d'utilisation de l'EIAH ; c'est la raison pour laquelle Lallé et al. ont développé un système d'assistance à la conception de techniques de diagnostic des connaissances qui permet de comparer plus facilement diverses techniques [10]. Néanmoins, il est admis que l'IRT a une capacité de prédiction généralement supérieure à celle du BKT [11].
Dans la situation pratique considérée, dans la mesure où le domaine et la population cibles ne sont pas précisés, il serait plus intéressant de choisir d'utiliser l'IRT sur ce critère de précision.

3.2 Esprit des modèles

Pour autant, on ne peut se contenter d'une comparaison de performance de précision sur une base de données. En effet, ces modèles peuvent être comparés selon d'autres critères d'utilisation, comme nous le montrons dans la suite.

Situation d'apprentissage. Une des hypothèses de l'IRT est que le niveau de connaissance de l'apprenant est fixe ; elle est questionnable puisqu'on peut se demander à quel moment un apprenant n'apprend plus, hors du moment où il maîtrise parfaitement le sujet. En dehors de cette remarque, on considère généralement que la connaissance de l'apprenant peut être considérée comme fixe lorsqu'il a déjà étudié et travaillé le sujet, qu'il est dans une situation d'évaluation, et lorsque les questions sont présentées à l'apprenant dans une fenêtre temporelle réduite. A l'inverse, dans le modèle du BKT la situation d'apprentissage supposée est celle d'une évolution d'un état de non maîtrise vers un état de maîtrise de la connaissance. Il est donc plus adapté dans une situation où l'apprenant débute et découvre le domaine.
Dans le cas pratique que nous considérons ici, il semble plus adapté d'utiliser le BKT.

Réalisme des modèles. Lorsque l'IRT est utilisé dans un outil durant une situation d'apprentissage, le modèle perd une grande partie de son réalisme et de son aspect explicatif (en opposition à son aspect prédictif). En effet, dans le cas de l'IRT, l'ordre des réponses de l'apprenant n'a pas d'importance : c'est à rebours de cette notion de progression et d'évolution qu'on a dans l'apprentissage.
Au contraire, le BKT inclut cette notion d'évolution de la connaissance. Dans un cadre évaluatif, comme une évaluation finale, il semble peu adapté : la probabilité de transition est alors théoriquement basse et beaucoup de réponses seront nécessaires pour faire évoluer les apprenants d'un état à l'autre. Il est donc tout à fait hors de propos de l'utiliser dans le cas d'une évaluation.
Cette tension entre un modèle « hautement complexe » qui a une bonne capacité de prédiction et un modèle « hautement structuré » qui a une bonne interprétabilité est mise en avant dans [12]. C'est également ce qu'il se joue ici entre l'IRT qui est plus prédictif qu'explicatif, et le BKT qui est plus explicatif que prédictif.

3.3 Conclusion du cas d'application

Dans le cas d'application considéré, nous cherchons à vérifier qu'un objectif pédagogique est atteint, autrement dit que le degré de maîtrise d'une connaissance est passée au-dessus d'un seuil. Il est donc nécessaire que l'estimation de cette maîtrise puisse être faite à la fois en posant le moins de questions possibles à l'apprenant (pour ne pas l'ennuyer) et à la fois avec le plus de certitude possible.
Il semble y avoir une contradiction entre le fait que l'IRT suppose une maîtrise fixe de l'apprenant, et le fait que l'IRT puisse prédire avec précision la justesse ou non de la réponse au moment où cette maîtrise varie le plus, autrement dit au moment des premières interactions avec la connaissance.
Dans des travaux futurs nous nous proposons donc d'étudier non pas uniquement la précision de ces deux algorithmes, mais l'évolution de cette précision dans la situation où l'apprenant découvre la connaissance : tend-on à avoir effectivement au début une prédominance du BKT qui s'efface petit à petit vers une prédominance de l'IRT ?
Si tel est le cas, il sera ensuite possible de combiner ces deux modèles en les utilisant en parallèle : il sera possible de choisir le meilleur en fonction du contexte, comme il est courant de faire lorsque l'on souhaite faire de la prédiction.

4 Conclusion

Dans cet article, nous avons comparé selon plusieurs angles les modèles de BKT et d'IRT ; ils pourraient également être comparés selon d'autres angles comme la problématique du départ à froid ou encore l'homogénéité de la population cible. Ces modèles reposent donc sur des hypothèses différentes qui ne les destinent pas au même usage. En cela, il peut ne pas être pertinent de les comparer sur des données similaires, ou uniquement sur des données qui pourraient « donner l'avantage » à l'un plutôt qu'à l'autre, car respectant mieux les hypothèses du modèle. Par ailleurs, lorsque l'on cherche à savoir lequel de ces modèles serait le plus pertinent à utiliser dans un outil lors de son développement, il est intéressant de regarder plus avant l'adéquation entre l'utilisation qui sera faite de l'outil et les hypothèses des modèles, afin de s'assurer que le calcul des paramètres donne des valeurs cohérentes.
Cette mise en miroir entre l'IRT et le BKT permet d'émettre l'hypothèse que le premier modèle est effectivement performant lorsque beaucoup de questions sont posées à l'apprenant, puisque la compréhension se stabilise autour d'un point petit à petit, là où le BKT est plus performant d'un point de vue de la prédiction sur les premières questions, et donc plus performant en général lorsque peu de questions sont posées à l'apprenant.
Nous nous proposons de vérifier cette hypothèse sur des ensembles de données dans le cas d'examens et dans le cas de plateformes d'apprentissage dans le cadre d'un travail futur.

Remerciements. Je souhaite remercier Matthieu Cisel pour son aide attentive lors de l'évolution de cette idée, mes directeurs de thèse Sebastian Ventura et Charles Tijus

pour leur accompagnement, ainsi que Benoit Praly et Simon Lemerle pour leurs relectures précieuses.

References

[1] B. Clement, D. Roy, P.-Y. Oudeyer, et M. Lopes, « Multi-armed bandits for intelligent tutoring systems », *J. Educ. Data Min.*, vol. 7, nº 2, p. 20-48, 2015.
[2] P. I. Pavlik Jr, H. Cen, et K. R. Koedinger, « Performance Factors Analysis–A New Alternative to Knowledge Tracing. », *Online Submiss.*, 2009.
[3] J.-J. Vie et H. Kashima, « Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing », *ArXiv181103388 Cs Stat*, nov. 2018.
[4] M. D. Reckase, *Multidimensional Item Response Theory*. New York, NY: Springer New York, 2009.
[5] G. L. Thorpe et A. Favia, « Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications. », 2012.
[6] M. V. Yudelson, K. R. Koedinger, et G. J. Gordon, « Individualized bayesian knowledge tracing models », in *International Conference on Artificial Intelligence in Education*, 2013, p. 171–180.
[7] R. S. J. d. Baker, A. T. Corbett, et V. Aleven, « More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing », in *Intelligent Tutoring Systems*, vol. 5091, B. P. Woolf, E. Aïmeur, R. Nkambou, et S. Lajoie, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, p. 406-415.
[8] Z. A. Pardos, N. T. Heffernan, C. Ruiz, et J. Beck, « Effective Skill Assessment Using Expectation Maximization in a Multi Network Temporal Bayesian Network », in *Proceedings of the The Young Researchers Track at the 9th International Conference on Intelligent Tutoring Systems*, 2008.
[9] S. Minn, Y. Yu, M. C. Desmarais, F. Zhu, et J. J. Vie, « Deep Knowledge Tracing and Dynamic Student Classification for Knowledge Tracing », in *arXiv:1809.08713 [cs]*, 2018.
[10] S. Lallé, V. Luengo, et N. Guin, « Assistance à la conception de techniques de diagnostic des connaissances », présenté à Environnements Informatiques pour l'Apprentissage Humain, Toulouse, 2013, p. 12.
[11] Z. A. Pardos et N. T. Heffernan, « KT-IDEM: introducing item difficulty to the knowledge tracing model », in *International Conference on User Modeling, Adaptation, and Personalization*, 2011, p. 243–254.
[12] M. Khajah, R. V. Lindsey, et M. C. Mozer, « How deep is knowledge tracing? », *ArXiv160402416 Cs*, mars 2016.