Évaluation des apprentissages, vers une approche sensible aux données pour l'estimation de la difficulté perçue d'items d'évaluation

Mohamed Lamgarraj

Évaluation des apprentissages, vers une approche sensible aux données pour l'estimation de la difficulté perçue d'items d'évaluation

Mohamed Lamgarraj[0009−0007−4708−1740], 2e année de thèse

Université de Picardie Jules Verne, Laboratoire MIS, 80000 Amiens, France
mohamed.lamgarraj@u-picardie.fr

Résumé Ces travaux de thèse s'intéressent à l'évaluation automatique des apprentissages, en particulier lors d'items d'évaluations de type Questions à Choix Multiples (QCM), en se focalisant sur la gestion de la difficulté et sa perception par l'apprenant. Cet article présente les méthodes traditionnelles d'estimation de difficulté d'items d'évaluation, suggérant l'utilisation de l'apprentissage automatique et du traitement du langage naturel, tout en notant leurs limites. Il présente la recherche, menée durant la première année de thèse, focalisée sur la création d'un jeu de données pour l'entraînement de modèles prédictifs, et explore la construction et l'entraînement de ces modèles. Il décrit le cadre pour une évaluation plus nuancée et prévoyant l'intégration de facteurs psychologiques dans la modélisation et les estimations futures de la difficulté.

Keywords: évaluation des apprentissages, traces, difficulté

1 Introduction

Dans l'enseignement supérieur l'évaluation des apprentissages joue un rôle déterminant dans la progression académique des étudiants. Dans ce cadre, l'estimation de la difficulté des questions de test (item) est cruciale pour garantir l'exactitude, voire l'équité, des évaluations des étudiants. La complexité de cette tâche découle de la diversité des méthodes d'évaluation qui peuvent être employées, ainsi que des facteurs affectant la perception de la difficulté d'une question ou d'une tâche. Ces derniers incluent non seulement le contenu de la question, sa structure (selon s nature), ou sa formulation, mais également les caractéristiques individuelles des étudiants évalués. Face à cette complexité, le recours à des outils numériques pour une évaluation automatique et plus objective des difficultés constitue une voie de recherche pour nos travaux, qui se focalisent dans cet article sur les items de type Questions à Choix Multiple (QCM).

Dans ce contexte la caractérisation de la difficulté items d'évaluation soulève une problématique complexe. Le présent article marque la première phase des travaux visant à cette caractérisation par la construction d'un modèle orienté vers l'estimation de cette difficulté à partir de traces, ainsi que les perspectives

d'intégration de caractéristiques psychométriques dans ce processus. Ce travail vise à examiner les approches et méthodologies utilisées dans la littérature pour prédire la difficulté d'items d'évaluation, ainsi que les ensembles de données mobilisés dans ces études. Partant de cette analyse, notre objectif est de développer un dataset conçu spécifiquement pour affiner les prédictions de difficulté via l'apprentissage automatique. La démarche présentée dans cet article aspire à améliorer la précision des modèles existants de prédiction de difficulté, et à établir les fondations pour une estimation plus sophistiquée et nuancée, intégrant des facteurs psychométriques plus complexes. Les sections suivantes présentent l'état de l'art, ainsi que la construction d'une première boucle d'analyse de données à partir d'un dataset de la littérature.

2 État de l'art

La psychologie, notamment dans les secteurs éducatif et cognitif, se penchent sur la difficulté à travers les prismes socio-cognitifs influençant la perception individuelle [2]. Elles analysent l'impact de facteurs tels que la mémoire de travail, la charge cognitive, et les stratégies métacognitives sur l'appréhension des tâches ardues. Les recherches évaluent comment les attributs des questions et les compétences des répondants modifient cette perception, en tenant compte des variations individuelles et émotionnelles [3]. Ces études reposent sur des méthodes heuristiques [4], et dépendent de l'intervention et de l'expertise humaine [1].

Différentes approches orientées données, sont également utilisées pour des tâches de prédiction de difficulté d'items de type QCM. La Théorie Classique des Tests (TCT) [5] introduit le concept d'indices de difficulté et de discrimination des items, offrant des informations sur la manière dont les items individuels du test distinguent les différents niveaux d'aptitude des individus. La Théorie de la Réponse à l'Item (TRI) [6] ou ses approches dérivées telles que le Modèle de Test Logistique Linéaire (LLTM) [8], vise à prédire la probabilité qu'un étudiant avec un niveau de compétence spécifique réponde correctement à un item. Des approches plus déterministes se basent sur les similarités entre la question, la réponse correcte et les distracteurs (réponses fausses d'un QCM). Enfin, certaines méthodes impliquent l'utilisation de modèles de langage pré-entraînés tels que BERT (Bidirectional Encoder Representations from Transformers) [7], pour capturer des informations contextuelles et peut être affiné pour des tâches spécifiques d'analyse du langage naturel pour prédire la difficulté des items.

Cette analyse bibliographique nous indique que nous pouvons nous reposer sur plusieurs approches analytiques. Si l'on souhaite se focaliser sur la mesure de la difficulté perçue, il est nécessaire de tenir compte du contexte de l'évaluation, ses enjeux, les conditions de passation et le temps alloué. La plupart des méthodes ci-dessus prennent peu en compte ces dimensions. Pour investir cette problématique, notre premier objectif est d'entraîner des modèles prédictifs sur des jeux de données issus de la littérature tenant compte du plus grand nombre possible de ces paramètres. Il s'agit d'évaluer la précision des modèles sur l'analyse et la prédiction de la difficulté. Ce premier jeu de données

permet de consolider les données pertinentes pour la difficulté perçue des items et ouvre la voir à l'adoption de méthodes de prédiction plus sophistiquées, et à l'intégration de facteurs psychologiques. La section suivante présente les données sur lesquelles nous avons travaillé.

3 Les données

Dans la littérature il existe divers ensembles de données relatifs à des évaluations en ligne, mais peu intègrent également des données sur les activités d'apprentissage ou satisfont le critère d'échelle nécessaire à l'entraînement des modèles. EdNet [9] est un ensemble de données hiérarchique, rassemblant deux ans de journaux d'interaction d'apprenants provenant du site Santa, conçu pour préparer au test TOEIC (Test of English for International Communication). Ces journaux totalisent 131 441 538 interactions collectées auprès de 784 309 étudiants. Ainsi EdNet répond au critère d'échelle, mais, pour constituer une composition de données capable de répondre à nos besoins, nous avons réalisé les tâches de regroupement, d'organisation et d'agrégation suivantes :

— Fusion des traces d'apprenants : EdNet (version-KT1) comporte 784 309 fichiers CSV (1 fichier par utilisateur) présentant les résultats des tests des apprenants. Cette tâche regroupe alors les traces en un premier fichier plat totalisant 95 293 926 lignes (et 6 colonnes).
— Intégration de données sur les questions : EdNet comportant également un kit 'Contents' de données relatives aux questions, ces dernières sont intégrées au fichier plat facilitant ainsi la corrélation des réponses.

Plusieurs attributs calculés et déduits, comme des ratios, des statistiques, ou d'autres métriques dérivées des données initiales, sont également ajoutées, soit pour une analyse plus fine des tendances dans le jeu de données, soit pour « aplatir » le fichier et ainsi focaliser les analyses sur les items. Ainsi, le jeu de données final apparaît alors sous deux formes :

Le dataset *utilisateur-question* contient des détails sur chaque question traitée par chaque apprenant (utilisateur) avec 35 colonnes et 79 929 968 entrées, Ce jeu de données est propice à l'ajout de paramètres liés à la difficulté perçue et l'intégration d'attributs liés à l'apprenant lui-même dans la prédiction et caractérisation de la difficulté. Le dataset *questions* se focalise sur les caractéristiques des questions dans leur globalité, avec 13 170 entrées représentant chaque question des tests et 27 colonnes issues de l'analyse de 8 687 observations par question en moyenne. Ce jeu de données présente l'agrégation de données du premier jeu.

La section suivante présente les expérimentations de modèles sur ces données.

4 Modèles et résultats

Forts de ces jeux de donnés, notre approche consiste à rechercher des modèles capables d'exploiter cet ensemble de données pour fournir une représentation et une estimation de la difficulté. Plusieurs algorithmes d'apprentissage automatique sont testés en vue de cette tache de prédiction et définir les caractéristiques

pertinentes. Ces expérimentations permettent également de mesurer la qualité et la pertinence des jeux de données.

Ainsi, les étiquettes de difficulté en apprentissage supervisé sont basées sur le taux de bonnes réponses par question. Une difficulté de 1 indique zéro réponse correcte (question difficile), et 0 signifie question facile. Trois classes de difficulté ont été créées : les questions avec une difficulté inférieure à 0.3 sont identifiées comme faciles (classe 0), celles dont la difficulté est comprise entre 0.3 et 0.6 sont considérées de difficulté moyenne (classe 1), et celles avec une difficulté supérieure à 0.6 sont jugées difficiles (classe 2).

L'expérimentation consiste à entraîner quatre modèles d'apprentissage automatique : régression logistique, arbre de décision, forêt aléatoire et XGBoost. Ces modèles ont été entraînés sur les deux ensembles de données, en utilisant 80% des données. Les 20% restants sont utilisés pour la validation des modèles, mettant en lumière les points forts et les faiblesses de chaque algorithme sur la prédiction de difficulté. Pour la deuxième version de notre ensemble de données, qui comprend de 79 929 968 lignes et 35 colonnes, seules 40 % des données ont été sélectionnées dans le voisinage du nombre moyen d'instances par question (8687 réponses). Pour évaluer les performances des modèles sur l'ensemble de données, le score F1 est utilisé. Ce dernier permet à la fois une bonne précision et un bon rappel (Recall) pour assurer une différenciation corrects entre les items difficiles et les items faciles. Les résultats sont présentés dans la table (Tab. 1).

Jeu de données	**Modèle**	**LR**	**DT**	**RF**	**XGB**
Orientées questions	Accuracy	0.54	0.91	0.91	0.93
	F1 score	0.50	0.85	0.81	0.82
Orientées utilisateurs-questions	Accuracy	-	0.80	0.89	0.74
	F1 score	-	0.65	0.70	0.71

TABLE 1 – Performance des modèles sur les deux jeux de données

La comparaison des performances des modèles sur les deux jeux de données a mis en évidence les subtilités de la prédiction de difficulté. Les modèles sont plus performants sur le jeu de données orienté questions. En effet, lorsque chaque question est unique cela facilitant la détection de motifs contrairement au jeu de données utilisateurs-questions, qui comporte des entrées multiples pour chaque question selon le nombre d'apprenants. La section suivante présente les perspectives de ces travaux de thèse.

5 Conclusion

Dans ce travail, nous nous sommes intéressés à la caractérisation de la difficulté d'items d'évaluation, avec comme objectif au cours de la thèse de s'orienter vers une représentation (modélisation) et une mesure de la difficulté perçue des items d'évaluation, en nous focalisant en premier lieu sur les items de type QCM. La construction d'un ensemble de données issues de la littérature contenant des informations propice à l'estimation de la difficulté a été présentée. Cet ensemble

de données a été érprouvé sur plusieurs prédictifs. Les résultats obtenus fournissent à la fois des informations sur la qualité des modèles et valide l'intérêt de notre premier ensemble de données pour caractériser la difficulté.

Une étude expérimentale en cours auprès d'étudiants en informatique et en psychologie, nous permet d'approfondir l'angle qualitatif des données pertinente pour cette prédiction. Elle vise à valider plusieurs facteurs humains influençant la difficulté perçue d'un item, et à expérimenter des dispositifs en mesure de capter certains de ces facteurs de perception au cours d'une évaluation. L'objectif est ainsi une amélioration et une expansion continues de l'ensemble de données en intégrant davantage de dimensions validée par la psychologie pour enrichir la caractérisation de la difficulté. Ce travail devrait conduire à la définition et à la mise en œuvre d'un "estimateur" de difficulté, améliorant notre compréhension de l'interaction complexe entre les apprenants et les composantes de l'évaluation. A plus long terme, il s'agira d'intégrer cette prédiction dans les algorithmes de génération automatique de tests d'évaluation à difficulté ciblée et de nous focaliser sur d'autres types d'items.

Références

1. Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. ETS Research Report Series, 2014(2), 1-8.
2. Beck, J., Stern, M., Woolf, B. P. (1997). Using the student model to control problem difficulty. In User Modeling : Proceedings of the Sixth International Conference UM97 Chia Laguna, 1997 (pp. 277-288). Springer.
3. Kubinger, K. D., Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats–An experiment in fundamental research on psychological assessment. Psychology science, 49(4), 361.
4. Lane, S., Raymond, M. R., Haladyna, T. M. (Eds.). (2015). Handbook of test development. Routledge.
5. Alagumalai, S., Curtis, D.D. (2005). *Classical Test Theory.* In : Maclean, R., et al. *Applied Rasch Measurement : A Book of Exemplars. Education in the Asia-Pacific Region : Issues, Concerns and Prospects*, vol 4. Springer, Dordrecht. https://doi.org/10.1007/1-4020-3076-21
6. Reckase, M. D. (1997). The past and future of multidimensional item response theory. Applied Psychological Measurement, 21(1), 25-36.
7. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. Preprint arXiv :1810.04805.
8. Poinstingl, H. (2009). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. Psychological Test and Assessment Modeling, 51(2), 123.
9. Choi, Y., et al.(2020). Ednet : A large-scale hierarchical dataset in education. In Artificial Intelligence in Education : 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21 (pp. 69-73). Springer.