Caractérisation des jeux de données

Auteur : Anne Boyer, Oriane Dermy, Inaya El Alaoui, Martin Lemaitre, Célina Treuillier
Nom de l'organisation : UL/Loria/CNRS
Date de création : 19-04-2022
Date de modification : 20-04-2022
Mots-cléfs :

Table des matières

1 Contexte

La plateforme LOLA propose des jeux de données différents, collectés dans des environnements d'apprentissage différents. Ils permettent notamment d'apprendre et de tester les différents algorithmes qui les exploitent. Il est donc essentiel de connaître leur représentativité en termes de caractéristiques des apprenants, afin de pouvoir éviter les biais d'apprentissage, d'identifier des apprenants caractéristiques des divers comportements numériques observés et de fournir des évaluations des résultats plus fines. C'est pourquoi la plateforme LOLA propose une caractérisation des jeux de données.

2 Principe

L'objectif est de fournir des outils de caractérisation des jeux de données. Le corpus est d'abord décrit en termes de contenu (traces disponibles) et de caractéristiques générales (données manquantes, ...) comme décrit dans la section 4.1. Cette description usuelle d'un jeu de données est ensuite complétée par la détermination des Personas présents dans le jeu de données. Un Persona peut être défini à partir de la définition donnée par [1] : "descriptions narratives d'apprenants typiques qui peuvent être identifiés à partir des centroïdes de méthodes de classification". Les Personas ainsi définis permettraient de représenter l'intégralité des étudiants présents dans le corpus et de les regrouper sous forme d'ensembles homogènes en termes de comportements numériques observés.

Le principe de détermination des Personas est schématisé dans la Figure 1. A partir des traces numériques disponibles dans le jeu de données, des indicateurs pédagogiques (cf. Section 4.2) sont déterminés automatiquement. Ces indicateurs pédagogiques vont servir à déterminer des classes d'utilisateurs, c'est-à-dire des groupes d'utilisateurs présentant des comportements numériques (observés au travers des traces) similaires. Ces classes sont ensuite traduites en Personas (cf. Section 5), qui donnent une description en langue naturelle du comportement type. Les indicateurs pédagogiques sont ensuite complétés par des indicateurs de comportement (cf. Section 4.3) qui pour chaque indicateur pédagogique et chaque classe décrivent sa dynamique au cours du temps. Ces nouveaux indicateurs sont ensuite traduits en langue naturelle pour compléter la description d'un Persona.

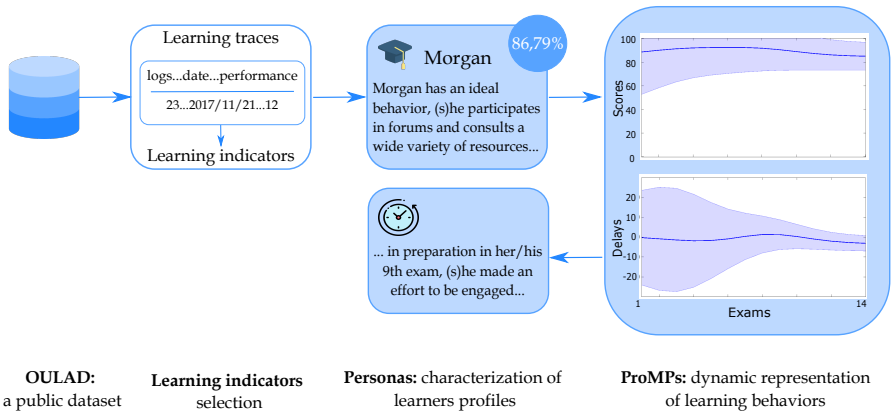

FIGURE 1 – Principe de la caractérisation d'un jeu de données

3 Description OULAD

Pour illustrer la caractérisation des jeux de données dans ce document, nous allons utiliser le jeu de données nommé “Open University Learning Analytics Dataset” (OULAD) [2]. Il rassemble des informations diverses à propos d'étudiants ayant suivi des cours à l'Université Ouverte du Royaume-Uni (Open University), l'une des plus grandes universités d'enseignement à distance au monde.

Le jeu de données OULAD est particulièrement utilisé et analysé car il est disponible gratuitement, et répond donc à une politique de reproductibilité des résultats. Il contient des données démographiques, des données d'interaction horodatées recueillies sur l'environnement numérique de travail (ENT) et les résultats des étudiants aux différentes évaluations. Il est important de préciser que les données sont totalement anonymisées.

OULAD contient des informations sur 32 593 étudiants ayant suivi des cours, appelés modules, en 2013 et 2014. Au total, le jeu de données contient des informations à propos de 22 modules différents, qui sont des cours de STEM (Sciences, Technologies, Ingénierie, Mathématiques) ou de Sciences Sociales. Ces cours peuvent être dispensés plusieurs fois durant l'année. Ils sont donc différenciés par l'année (2013 ou 2014), ainsi que le mois durant lequel ils commencent. Par exemple, un module dispensé en février 2013 porterait la mention 2013B. Chaque étudiant peut avoir quatre résultats différents : réussite, félicitations, échec ou abandon.

Nous considérons les données correspondant au module de STEM appelé D, commençant en février 2013 (référence : 2013B). Il a duré 240 jours, inclut 14 évaluations pour 1303 étudiants. La table 1 précise pour chacun des quatre résultats possibles le nombre d'étudiants concernés.

Résultat	**Nombre étudiants**	**% du corpus initial**
Succès	456	35,0%
Échec	361	27,7%
Abandon	482	33,2%
Félicitations	54	4,1%

Table 1 – Caractéristiques du jeu de données OULAD.

Les informations contenues dans OULAD sont multiples et détaillées : nous nous concentrons principalement sur l'activité des apprenants sur l'ENT (données d'interaction) et les modalités de rendu et les performances aux examens. Parmi les données démographiques disponibles, seule l'information de genre pourra être utilisée.

4 Indicateurs

4.1 Caractérisations des données

Afin de pouvoir travailler sereinement avec des données, il est important de s'assurer que le jeu de données est de bonne qualité. Pour cela, la méthode préférable est de définir des indicateurs de qualité qui permettent d'évaluer un jeu de données. Chaque indicateur doit permettre de produire un score et la moyenne des scores correspond au score global du jeu de données. Ainsi, on obtient un score de qualité objectif qui permet notamment de comparer plusieurs jeux de données.

Il n'existe pas de liste d'indicateurs de qualité universels applicables à n'importe quel jeu de données pour en estimer la qualité, il faut donc choisir des indicateurs spécifiquement pour un jeu de données. Pour être exploitables, les indicateurs définis doivent répondre à plusieurs contraintes [3] :

- Mesurabilité : les indicateurs doivent produire une donnée numérique, bornée et normalisée. C'est à dire que chaque indicateur doit retourner une valeur comprise entre 0 et 1 pour permettre de calculer une moyenne des indicateurs et obtenir un score de qualité global.
- Interprétabilité : Les indicateurs doivent être compréhensibles pour les personnes chargées d'interpréter les résultats.
- Agrégation : Si un indicateur est appliqué au niveau d'une valeur ou d'une ligne, il faut pouvoir regrouper ses résultats pour obtenir un score qui s'applique au jeu de données complet.
- Faisabilité : Les indicateurs doivent être applicables au jeu de données, donc calculables à partir des entrées disponibles et préférablement de manière automatique.

Pour définir des indicateurs pertinents, il faut d'abord identifier des caractéristiques de qualité. Pour chaque caractéristique il s'agit ensuite de trouver un ou plusieurs indicateurs qui entrent dans la définition de cette caractéristique tout en répondant aux contraintes précédemment citées. Ces caractéristiques dépendent également du jeu de données même s'il existe des caractéristiques communes à un grand nombre de jeux de données. Voici quelques exemples de caractéristiques avec des indicateurs associés [4] :

- Complétude : la complétude correspond à l'estimation du nombre de données manquantes.
  Exemples d'indicateurs : pourcentage de données renseignées, pourcentage de lignes dont toutes les données sont renseignées.
- Précision : la précision correspond à l'estimation du nombre de données correctes.
  Exemple : pourcentage de données qui correspondent à leur contraintes (type de données, format, valeurs minimales et maximales, ...)
- Actualité : l'actualité correspond à l'estimation du nombre de données qui sont à jour.
  Exemple : pourcentage de données à jour, délais entre la récolte
- Compréhensibilité : la compréhensibilité correspond à la facilité avec laquelle les données peuvent être comprises par les utilisateurs.
  Exemples : pourcentage de colonnes qui possèdent des métadonnées, pourcentage de données qui sont dans un format compréhensible.

4.2 Indicateurs pédagogiques et indicateurs de comportement dynamique

4.2.1 Définition

Iksal [5] définit un indicateur de la manière suivante : "Un indicateur est un observable signifiant sur le plan pédagogique, calculé ou établi à l'aide d'observés, et témoignant de la qualité de l'interaction, de l'activité et de l'apprentissage dans un Environnement Informatique pour l'Apprentissage Humain (EIAH). Il est défini en fonction d'un objectif d'observation et motivé par un objectif pédagogique". Dans la suite de ce document, nous appellerons indicateur pédagogique un indicateur répondant à cette définition. Dans notre contexte, le choix et le calcul des indicateurs pédagogiques dépendent des données disponibles et par conséquent du jeu de données utilisé. Dans le cadre du jeu de données OULAD, il est important de noter que les données concentrent des informations sur 20 types de ressources pédagogiques, avec lesquels les utilisateurs peuvent interagir. Cependant, certains types d'activités ont plus d'influence sur les résultats de l'apprentissage : les activités de forum, contenu relatif au cours, page d'accueil et sous-page (telles qu'intitulées dans OULAD) sont, par exemple, les prédicteurs les plus importants de l'engagement selon Hussain et al.[6]. Nous n'avons donc retenu que ces quatre types d'activités.

4.2.2 Description des indicateurs pédagogiques

Les indicateurs pédagogiques doivent être évalués automatiquement à partir des traces d'apprentissage disponibles. Les indicateurs suivants, ont tous été évalués dans la littérature à partir de traces d'apprentissage et seront proposés pour la caractérisation des jeux de données, lorsque les traces numériques nécessaires à leur détermination sont présentes :

- **Engagement :** adoption d'un comportement engagé par l'étudiant envers l'activité d'apprentissage [6].
- **Performance :** différents résultats obtenus par l'étudiant [7].
- **Réactivité :** réponse à temps pour les différents évènements liés au cours [8].
- **Régularité :** investissement dans la tâche d'apprentissage à intervalles de temps réguliers et rapprochés [8].
- **Curiosité :** motivation intrinsèque de l'étudiant [9].

Reprenons l'exemple du jeu de données OULAD (décrit dans la Section 3). Il est possible de déterminer automatiquement chacun des cinq indicateurs pédagogiques à partir des traces disponibles. La table 2 résume les différents modes de calcul des valeurs qui caractérisent dans ce cas chaque indicateur pédagogique [10].

Indicateur	**Détermination de l'indicateur dans le contexte d'OULAD**
Engagement	— nombre de clics total sur toute la durée d'un cours — nombre de clics par activité
Performance	— les 14 notes
Réactivité	— différence entre la date de rendu de l'élève et la date de rendu limite
Régularité	— nombre de jours actifs par type d'activité — nombre total de jours actifs — moyenne de clics par jour sur différentes activités — total des clics toutes activités confondues
Curiosité	— nombre de types d'activités différentes consultées — nombre de ressources différentes consultées

Table 2 – Détermination des indicateurs pédagogiques dans OULAD.

4.3 Indicateur de comportement dynamique

Afin de caractériser le comportement des utilisateurs correspondant à un jeu de données, nous avons inventé un nouvel indicateur [11], nommé indicateur de comportement dynamique. Un exemple sur le jeu de données OULAD est présenté Figure 2. Il s'agit d'un indicateur visuel, facile à interpréter, qui modélise les changements comportementaux des utilisateurs à partir de motifs récupérés à l'aide de fouille de données sur des données qui peuvent être pauvres et génériques. Cette approche a notamment l'avantage de représenter l'évolution globale du comportement des utilisateurs, sans donner d'informations spécifiques concernant ces utilisateurs.

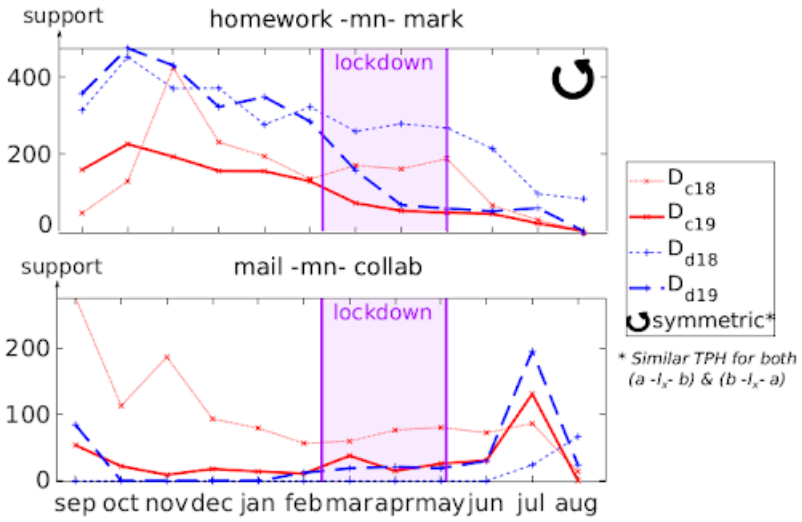

FIGURE 2 – Exemple d'indicateur de comportement numérique sur des données de confinement.

À partir de motifs temporels détectés à l'aide de fouille de données, nous représentons graphiquement l'évolution de ces motifs dans le temps. Nous avons donc un histogramme de motifs temporels. Cette méthode a été appliquée à des données correspondant à différents services (services de mails, de devoirs, de notes...) des ENT utilisés par des étudiants provenant de différentes zones françaises (une ville et un département). En appliquant cette méthode au cours de l'année pré-covid et l'année de début d'épidémie Covid-19, nous avons ainsi pu repérer un changement de dynamique profond chez les étudiants lors du premier confinement.

Cette méthode est en cours d'intégration dans la plateforme LOLA.

5 Détermination des Personnas

Les groupes homogènes d'étudiants sont identifiés grâce aux indicateurs pédagogiques. Les Personas sont couramment utilisés lors de la phase de développement de services numériques, notamment en UX design [12] : ils représentent des utilisateurs types, auxquels le service doit répondre. Dans notre cas, l'objectif des personas d'apprenants est différent : ils permettent de caractériser la représentativité d'un jeu de données d'une manière facilement compréhensible et interprétable pour un expert pédagogique ou un chercheur.

Dans notre contexte, chaque Persona présente une description narrative d'un étudiant fictif : il contient quelques informations démographiques (nom, sexe et âge), associées à une description textuelle donnant des indices essentiels sur le comportement d'apprentissage, selon les indicateurs d'apprentissage. Cette description permet d'incarner les résultats renvoyés par le processus de classification : toute personne susceptible de la lire peut la comprendre facilement. La figure 3 présente deux exemples de Personas déterminés sur le jeu de données OULAD. La fréquence (% donné dans le Persona) permet de donner l'importance du Persona dans le jeu de données.

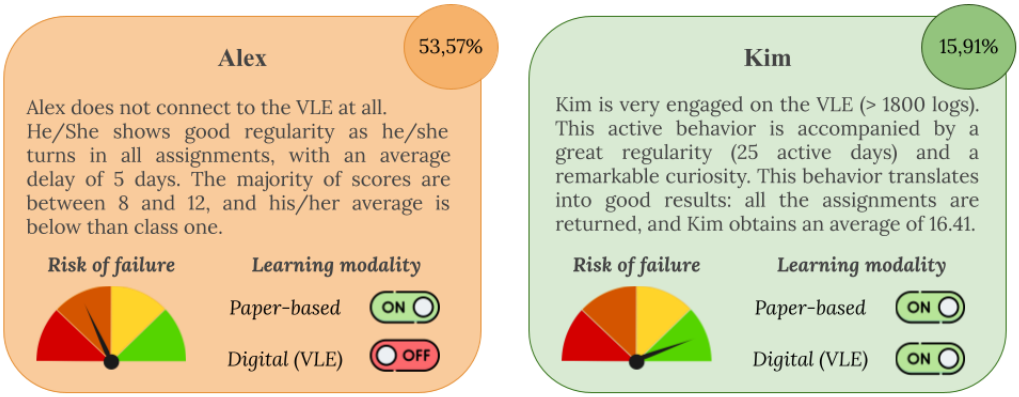

FIGURE 3 – Deux exemples de Personas

Les prénoms choisis sont volontairement non genrés. Des renseignements complémentaires pourront être trouvés dans [10]. Chaque Persona est ensuite complété par la traduction en langue naturelle des indicateurs de comportement dynamique calculés pour chaque indicateur pédagogique. La figure 4 illustre un exemple de Persona ainsi déterminé [11].

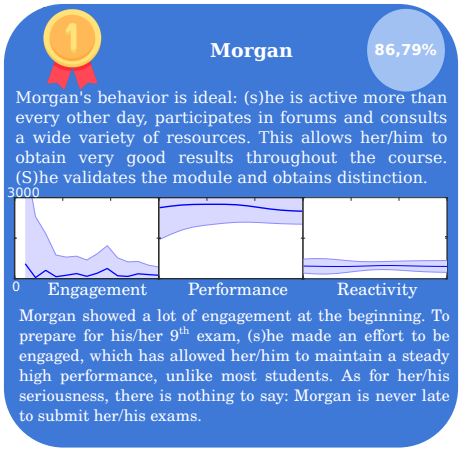

FIGURE 4 – Exemple de Persona complété

6 Conclusion

Cette méthode permet de fournir une description plus complète des jeux de donnés proposés dans la plateforme LOLA et ainsi de les caractériser en termes facilement compréhensibles par tout un chacun. Les Personas permettent également une évaluation plus fine, puisqu'il sera possible de déterminer des "apprenants mystères" correspondant à chaque Persona (y compris les Personas correspondant à des profils atypiques), de calculer les indicateurs de performances et d'évaluation d'une manière générale par Persona ...

Références

[1] C. Brooks and J. Greer, "Explaining predictive models to learning specialists using personas," in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, pp. 26–30, ACM.

[2] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," vol. 4, no. 1, p. 170171.

[3] B. Heinrich, M. Kaiser, and M. Klier, "How to measure data quality ? a metric based approach," p. 15.

[4] A. Vetro, L. Canova, M. Torchiano, C. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework : Definition and application to open government data," *Government Information Quarterly*, vol. 33, 02 2016.

[5] S. Iksal, "Ingénierie de l'observation basée sur la prescription en EIAH," p. 128.

[6] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," vol. 2018, pp. 1–21.

[7] K. E. Arnold and M. D. Pistilli, "Course signals at purdue : using learning analytics to increase student success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 267–270, ACM.

[8] M. S. Boroujeni, K. Sharma, L. Kidzinski, and P. Dillenbourg, "How to quantify student's regularity ?," p. 14.

[9] G. Pluck and H. L. Johnson, "Stimulating curiosity to enhance learning," p. 9.

[10] C. Treuillier and A. Boyer, "Identification of class-representative learner personas," p. 8.

[11] A. Dermy, 0. ; Boyer and A. Roussanaly, "A dynamic indicator to model students' digital behavior,"

[12] C. Lallemand and G. Gronier, *Méthodes de design UX : 30 méthodes fondamentales pour concevoir et évaluer les systèmes interactifs*. Design Web, Eyrolles.