L'apprentissage automatique comme base du suivi d'élèves et de l'amélioration de formations

Angela Bovo, Stephane Sanchez, Olivier Héguy, Yves Duthen

L'apprentissage automatique comme base du suivi d'élèves et de l'amélioration de formations

Angela Bovo[1,2], Stéphane Sanchez[1], Olivier Héguy[2], and Yves Duthen[1]

[1] IRIT - Université Toulouse 1
firstname.lastname@univ-tlse1.fr
[2] Andil
firstname.lastname@andil.fr

Abstract. Cet article vise à présenter un projet de recherche dont le but est d'utiliser des méthodes d'intelligence artificielle et de fouille de données pour l'e-learning. Nous proposons des solutions techniques au problème du suivi des élèves en formation et de l'amélioration des formations proposées. Notre solution prend la forme d'une application qui centralisera des données issues de LMS et permettra de les examiner et de les analyser en utilisant des méthodes de l'intelligence artificielle. Cette application pourra dans un deuxième temps servir de base à la création d'un tuteur virtuel intelligent. Nous détaillons nos propositions concernant les méthodes à employer, l'architecture de l'application et les éléments choisis pour servir d'indicateurs et d'attributs d'apprentissage automatique, et analysons les résultats préliminaires d'un partitionnement de données.

Keywords: prédiction, extraction de connaissances, apprentissage automatique, tuteur virtuel intelligent

1 Notre projet de recherche

Ce projet s'inscrit dans le cadre d'une thèse CIFRE débutée en février 2011 entre l'IRIT (Institut de Recherche en Informatique de Toulouse) et l'entreprise Andil, société informatique spécialiste des technologies de l'information et de la formation, qui travaille dans l'e-learning. Le but du projet est d'utiliser les méthodes, outils et technologies de l'informatique, de la statistique et de l'intelligence artificielle pour donner aux pédagogues de meilleurs outils de suivi des élèves en formation e-learning et d'ingénierie des formations. Par ailleurs, nous travaillons sur des données réelles issues d'un partenariat avec l'institut de formation professionnelle Juriscampus.

Notre premier objectif avec ce projet est d'améliorer les possibilités de suivi des élèves en cours de formation pour éviter leur décrochage ou leur échec. Nous pensons que si certaines situations d'échec sont peut-être inéluctables (circonstances personnelles, mauvaise adaptation des désirs de l'apprenant à la formation proposée), une bonne partie d'entre elles pourraient être évitées par un repérage précoce qui donnerait lieu à un recadrage et un suivi plus attentif et personnalisé. Il nous paraît important d'agir le plus tôt possible face à un élève

qui commence à décrocher pour qu'il ne prenne pas trop de retard par rapport aux autres. Une fois un retard détecté, un enseignant pourrait contacter l'élève pour vérifier s'il s'agit d'un décrochage ou juste d'un rythme plus lent mais pas nécessairement grave.

Pour ce faire, nous pensons nous baser notamment sur l'étude des traces d'activité modélisées, comme dans [1] et [2]. Or, à ce jour, le LMS que nous utilisons le plus couramment, Moodle [3, 4], ne dispose que de quelques statistiques difficiles à examiner et analyser. Notamment, ces statistiques sont uniquement individuelles, ce qui empêche d'avoir une vision globale d'un groupe d'élèves et de pouvoir les comparer entre eux. En outre, il ne permet que d'avoir accès à une suite de logs d'activité sans guère proposer de vision plus synthétique, hormis quelques graphiques pour les données de connexion. Nous avons discuté avec une responsable du suivi des élèves qui nous a dit que, pour cette raison, elle s'en tenait à deux seuls indicateurs du succès d'un étudiant : la date de sa dernière connexion et le nombre d'activités notées qu'il avait réalisées. Cela nous semble dommage en regard des vastes possibilités offertes par l'enregistrement que Moodle fait de chaque clic d'un apprenant.

Notre deuxième objectif est d'améliorer les formations mises à disposition des élèves. Il serait très intéressant de repérer qu'un quiz est plus difficile que prévu ou qu'une leçon ne retient pas l'intérêt des étudiants qui ne la lisent pas jusqu'au bout afin de pouvoir les améliorer en conséquence. Nous souhaitons également évaluer de manière plus précise l'importance des connaissances pré-requises à la formation et de l'ordre de parcours des ressources pendant la formation. Idéalement, cela nous servirait à extraire les parcours suivis par les meilleurs élèves de chaque formation et voir s'ils diffèrent de ceux suivis par les moins performants pour juger de l'importance de cet ordre [5].

À côté de ces objectifs primaires, nous avons également des buts qui nous semblent importants à atteindre pour faciliter la réalisation de ces objectifs.

Nous nous sommes fixé comme premier but secondaire de centraliser les données fournies par une formation en e-learning. Actuellement, beaucoup de données sont dispersées en deux principaux endroits : les traces d'activités sont présentes sur le LMS (Moodle ici) et d'autres données (administratives, d'activité en présentiel, d'historique de contact et de communication, de notes aux examens finaux) sont conservées par les responsables de formation et leur équipe administrative ainsi que par les divers intervenants et enseignants, parfois dans des solutions peu adaptées comme des tableurs Excel. Nous voulons dons rassembler, autant que possible, les données pédagogiques pertinentes en un seul endroit de stockage. Cette centralisation nous permettra également de pouvoir comparer plusieurs types de LMS, en s'affranchissant de leur format spécifique de stockage.

Notre deuxième but secondaire vise à automatiser le plus de tâches possibles parmi celles qui sont actuellement faites par des humains comme les responsables de formation. Dans encore trop de cas, les personnes que nous avons contactées devaient jusqu'à présent recueillir et traiter les données manuellement : recopier dans un tableur les données fournies par l'outil de statistiques de Moodle pour avoir une vue d'ensemble, repérer les élèves en difficulté, interroger ou soumettre

des questionnaires aux élèves quant à la qualité des formations, interroger les enseignants pour avoir leur avis sur les élèves, envoyer des mails et téléphoner aux élèves et aux enseignants pour leur fournir des retours...

Notre troisième objectif secondaire serait de pouvoir tester des algorithmes innovants, non encore employés dans le domaine des EIAH, afin de comparer leur efficacité à celle des méthodes déjà testées et éprouvées.

Pour atteindre ces buts, nous avons proposé une architecture de recueil et traitement des données issues de formation e-learning. Nous avons déjà commencé à réaliser les applications correspondantes et à mener des premiers tests dessus.

Nous avons également déjà établi une réflexion sur le choix d'attributs d'apprentissage qui nous semblent à la fois pertinents à notre projet et suffisamment génériques pour pouvoir être utilisés par d'autres.

Nous avons pu mener de premières expériences de partitionnement de données sur des données issues de formations réelles, grâce à notre institut de formation partenaire, dont nous présentons les résultats accompagnés d'une analyse.

La suite de l'article est organisée comme suit : la section 2 contient nos propositions quant aux méthodes à employer pour parvenir à nos fins, à l'architecture qui permettra de les réunir ainsi qu'aux attributs d'apprentissage que nous allons utiliser ; la section 3 présente l'état d'avancement de notre projet ; la section 4 présente la méthode que nous avons utilisée pour mener notre première expérience, dont les résultats sont expliqués dans la section 5 ; enfin, la section 6 propose une conclusion et des perspectives sur notre projet.

2 Propositions

2.1 Méthodes employées

Pour améliorer le suivi des élèves, nous comptons avoir recours à plusieurs méthodes. La première est de disposer de statistiques fiables sur les actions des étudiants. Nous avons besoin de savoir s'ils se connectent souvent à la plateforme, s'ils lisent les leçons, font les quiz, participent aux discussions sur les forums, etc. Nous avons besoin de pouvoir consulter ces statistiques à plusieurs niveaux de granularité différent, pour un élève seul ou pour un groupe d'élèves. Au niveau de granularité le plus bas, nous souhaitons pouvoir montrer chaque instance d'un type d'action (par exemple lire une leçon) avec la date associée. Au niveau de granularité le plus haut, nous souhaiterions donner à chaque élève une note sur 10 représentative de la quantité et de la qualité de son activité d'apprentissage. Nous avons baptisé cette note l'Indice général d'apprentissage (IGA). Aux niveaux intermédiaires se trouveraient diverses statistiques et des indicateurs. Nous avons également besoin que ces indications soient très visuelles, avec des graphiques pertinents proposés et une colorimétrie permettant de détecter au premier coup d'œil si un élève est en difficulté par rapport à ses collègues.

La deuxième est d'utiliser des méthodes d'intelligence artificielle pour traiter ces statistiques de façon plus fouillée. Dans un premier temps, nous allons

utiliser des méthodes de type partitionnement de données [6, 7] pour détecter des groupes d'élèves de comportement semblable [8–11]. Il est important de repérer les différentes méthodes d'apprentissage mises en œuvre par les étudiants pour pouvoir comprendre leurs spécificités et pouvoir leur proposer une offre de parcours adaptée. Dans un deuxième temps, la classification automatique des élèves permettra d'essayer de comprendre ce qui fait la différence entre ceux qui réussissent et ceux qui échouent [11]. Enfin, la régression permettra d'essayer de prédire la note qu'ils obtiendront à l'examen final. Nous allons appliquer les connaissances obtenues sur les données des formations passées pour faire de la prédiction [12] sur les données des formations en cours. Si nous repérons qu'historiquement, un type de comportement a conduit à un échec ou un abandon de la formation, il est important d'agir vite et d'en informer l'apprenant.

Dans un premier temps, nous allons choisir pour cela divers algorithmes utilisés classiquement dans ce domaine. Mais pour répondre à notre troisième objectif secondaire, qui est de tester d'autres méthodes peu utilisées dans ce domaine, nous avons choisi un algorithme encore peu testé, mais qui nous semble adapté à nos données. Cet algorithme, HTM (Hierarchical Temporal Memory), dans sa version Cortical Learning Algorithm [13], est un algorithme de représentation de données inspiré du fonctionnement du néocortex humain. Selon ses concepteurs, cet algorithme est adapté aux données non aléatoires mais issues de causes sous-jacentes spatiales et temporelles. Il peut ensuite être utilisé pour avoir en sortie de la classification ou de la prédiction. Les auteurs n'ont pour l'instant décrit que le cœur de leur algorithme, sans ces sorties. Les données de logs de Moodle nous semblent répondre parfaitement à cette spécification. En conséquence, nous souhaitons implémenter l'algorithme CLA avec des sorties de classification et de prédiction pour observer ses résultats sur nos données.

Pour améliorer les formations mises à disposition des élèves, nous allons reproduire cette méthodologie qui va consister à mettre en place dans un premier temps des indicateurs basés sur des statistiques, puis à traiter ces indicateurs par des méthodes d'intelligence artificielle dans les cas pertinents. Une fois que nous disposerons de profils d'apprentissage, nous pourrons même proposer un parcours adapté au style d'apprentissage de l'élève.

Afin de centraliser les données d'apprentissage, nous allons regrouper les données issues des logs de LMS et les autres données dans une seule base de données. Notre application de suivi d'élèves et d'amélioration de formation devra idéalement n'utiliser que cette base comme source de données. Nous proposerons en conséquence des méthodes d'import de données existantes et de rajout manuel de données au fil de l'eau. Pour l'instant, nous nous basons sur des traces enregistrées par une plate-forme Moodle, mais les indicateurs que nous en tirons sont suffisamment génériques pour pouvoir être élargis à l'avenir à d'autres LMS comme sources de données.

Pour atteindre notre objectif d'automatisation du travail du responsable de formation, dans un premier temps, nous souhaitons automatiser la collecte de données sur les étudiants, ce qui sera fait en même temps que le regroupement de ces données, par des méthodes d'import automatique.

Dans un deuxième temps, nous pensons également automatiser une partie du retour de conseils envers les étudiants dans certaines situations pré-définies. Par exemple, il pourrait être intéressant de suggérer à un élève qui vient de finir de lire une leçon de poursuivre par le quiz qui lui est associé, ou inversement, de recommander à un élève qui n'aura pas brillé à un quiz de retourner lire le chapitre correspondant. On pourrait également envoyer automatiquement des e-mails à un étudiant peu présent pour l'inviter à se connecter plus fréquemment ou le prévenir de l'ouverture d'une nouvelle activité. Cette automatisation pourrait aussi prendre la forme d'un tuteur virtuel intelligent [14] évoluant et agissant sur le LMS en interaction directe avec les élèves et leurs formateurs.

Nous n'avons pour l'instant pas prévu d'approche réflexive visant à donner directement à l'étudiant accès à son positionnement pour qu'il choisisse lui-même comment réagir, mais cela pourra être étudié si jugé pertinent par les responsables de formation. Cela peut poser problème dans le sens où, comme beaucoup de nos données sont relatives à un positionnement dans le groupe, l'étudiant aurait alors nécessairement accès à beaucoup de données concernant ses camarades, auxquelles ces derniers n'auraient pas forcément envie de lui donner accès.

2.2 Architecture proposée

Nous proposons d'intégrer les différentes méthodes et solutions proposées en une seule application Web. Nous sommes actuellement en train de la développer. Pour cela, nous avons fait le choix d'utiliser le langage Java et les frameworks Wicket, Hibernate, Spring et Shiro. Toutes nos données seront importées dans une base de données MySQL. Pour ce qui est des algorithmes d'apprentissage automatique, nous avons fait le choix d'utiliser les implémentations fournies par la bibliothèque libre Weka [15].

Nous importerons régulièrement depuis Moodle toutes les données pertinentes dans notre base de données : formations, cours, leçons, ressources, activités, logs et notes. Ces données seront ensuite traitées de la façon décrite dans la figure 1, où elles passeront par les stades de traces brutes, traces modélisées, et indicateurs.

2.3 Le choix de nos attributs d'apprentissage

Nous avons défini en collaboration avec des responsables de formation des indicateurs [16] qui ont deux usages dans notre application. Le premier est l'affichage à destination d'un humain, qui pourra associer cette note au niveau de l'élève dans le domaine en question. Le deuxième est l'utilisation en tant qu'attribut pour les divers algorithmes d'apprentissage automatique que nous allons employer.

Pour l'instant, nous n'avons implémenté que les indicateurs correspondant aux données issues des logs de Moodle, et nous y rajouterons progressivement les indicateurs venant d'autres sources au fur et à mesure que nous intégrons

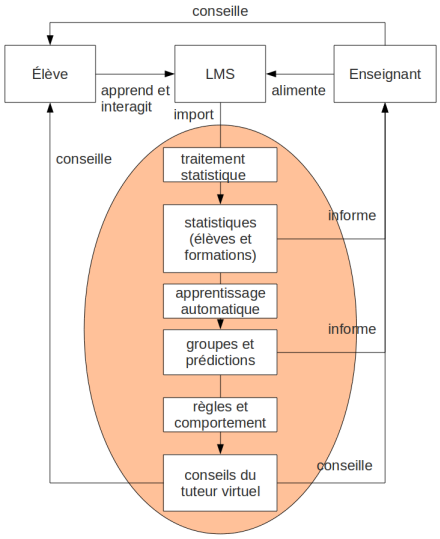


Fig. 1. L'architecture proposée

ces autres sources à notre application. Certains de ces indicateurs ne sont justement qu'indicatifs, car ils ne peuvent être calculés avec exactitude (par exemple le temps passé en ligne), mais ils peuvent néanmoins apporter un surcroît d'information.

Les indicateurs que nous avons définis sont présentés dans la figure 2. Nous voyons qu'ils s'agrègent naturellement en grands thèmes dont la majorité concerne une quantité d'activité et un seul, les notes, vérifie que cette activité porte ses fruits. Nous avons essayé de refléter aussi bien les aspects directement liés à l'apprentissage que les aspects plus sociaux qui peuvent jouer un rôle important dans la motivation des étudiants. Nous regarderons ultérieurement comment modéliser nos indicateurs selon le langage UTL [2].

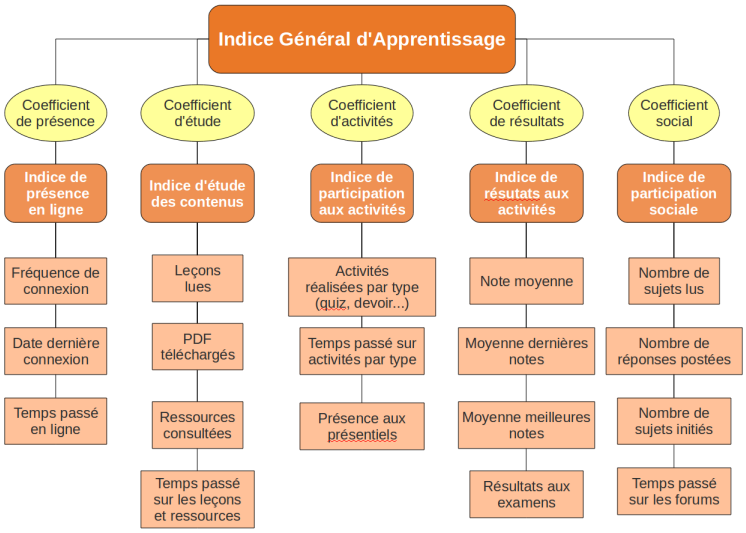


Fig. 2. Les indicateurs proposés

3 Avancement

L'application la plus importante au regard de notre projet, baptisée GIGA (Gestionnaire d'indices généraux d'apprentissage), est déjà partiellement réalisée. Elle comprend déjà toutes les fonctionnalités d'import de données depuis Moodle de production et traitement de statistiques aux fins de suivi des élèves. Elle est déjà

en usage auprès d'un institut de formation dans sa version actuelle et poursuit sa commercialisation. En parallèle, les parties de retour sur la qualité des formations, d'ajout de données issues d'autre sources et d'apprentissage automatique sont encore en cours de développement.

Nous avons également réalisé des premiers tests de partitionnement de données que nous présentons ci-dessous.

4 Méthodologie de test

Comme mentionné plus haut, nous avons utilisé Weka, pour son API Java facile à intégrer avec notre application et la variété de méthodes qu'il propose. Dans un premier temps, nous avons converti nos données sous la forme d'une liste de notes pour chaque attribut mentionné plus haut. Nous avons ensuite transformé ces données en attributs et instances au format Weka. Les algorithmes de partitionnement de données de Weka prennent ces instances en entrée et renvoient en sortie des groupes de ces instances, que nous convertissons ensuite en groupes d'élèves. Ensuite, pour analyser sémantiquement ces groupes, nous examinons les indicateurs pour chaque groupe d'élèves.

Pour tester la justesse de nos groupes, nous utilisons la contre-validation. Nous avons également pris une moyenne des résultats obtenus sur plusieurs simulations avec des graines aléatoires différentes. Nous avons choisi les algorithmes suivants proposés par Weka: Expectation Maximisation, Hierarchical Clustering, Simple K-Means et X-Means. Parmi ces algorithmes, X-Means est à mettre à part puisqu'il détermine de lui-même le nombre de groupes qui lui semble le plus pertinent. Les trois autres prennent en paramètre le nombre de groupes souhaité. En conséquence, pour ceux-là (K-Means, EM et Hierarchical Clustering), nous avons fait plusieurs simulations en faisant varier ce paramètre, afin de voir quelle valeur donnait les résultats correspondant le mieux à la structure naturelle des données. Au vu de nos données, nous avons choisi des valeurs allant de 2 à 5, qui correspondent aux valeurs renvoyées par X-Means, ce qui valide ce choix.

Nous avons choisi 3 formations différentes, dont deux promotions différentes d'une même formation, que nous appellerons formations A1 et A2, et une formation différente B. La formation A1 a 56 élèves, A2 en a 15 et B en a 30. A1 et A2 durent environ un an, tandis que B dure trois mois.

5 Résultats

5.1 Nombre optimal de groupes

Les figures suivantes montrent les résultats des quatre algorithmes sur nos trois jeux de données. Le premier montre le nombre moyen de groupes proposé par X-Means pour décrire au mieux nos données. Les trois autres graphes montrent l'erreur moyenne obtenue en fonction du nombre de groupes fixé pour les algorithmes K-Means, Hierarchical clustering et Expectation Maximisation.

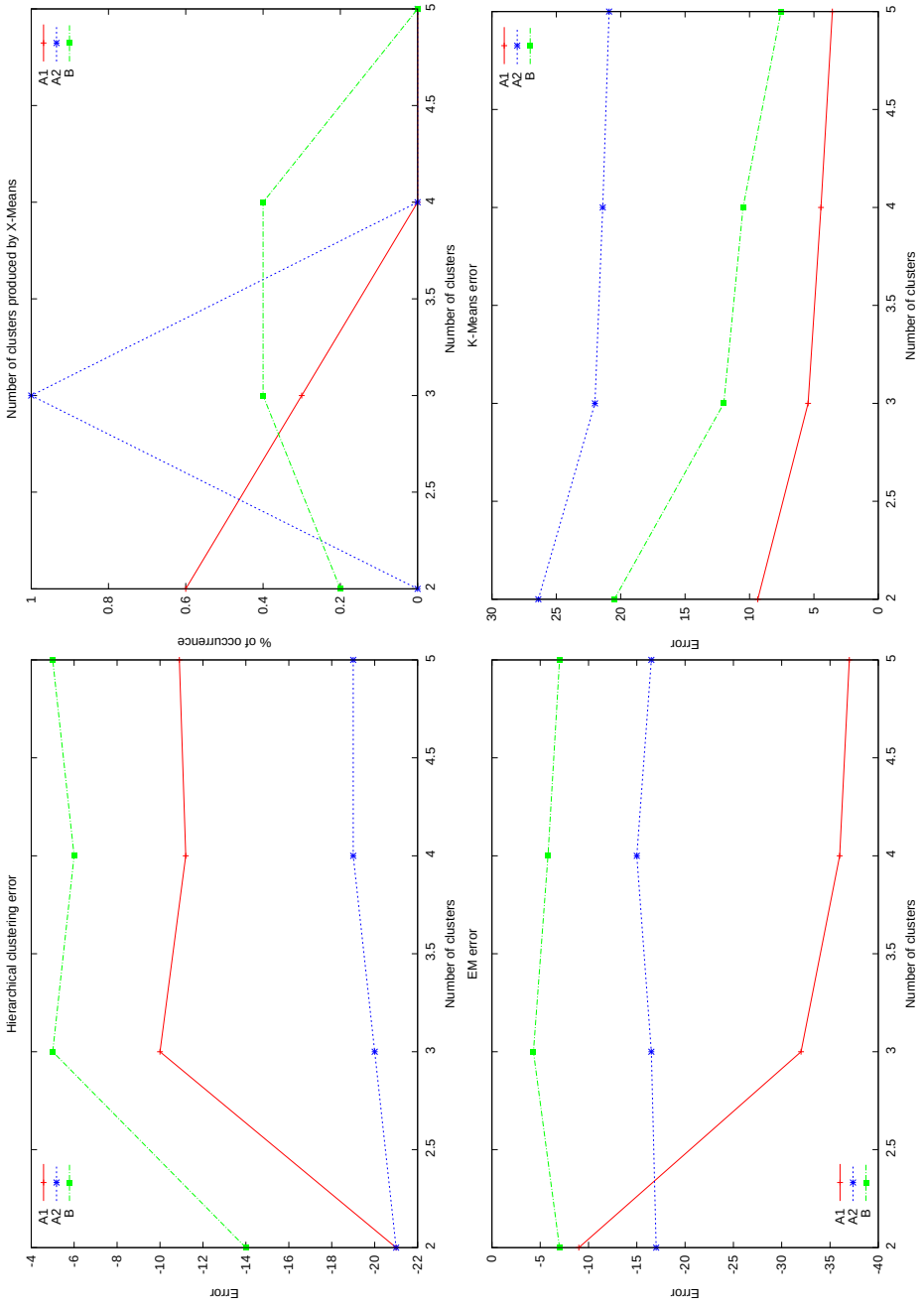


Fig. 3. Nombre optimal de groupes trouvé par chaque algorithme pour nos trois formations de test

Pour la formation A1 (courbe rouge en trait plein), X-Means propose 2 groupes, ce qui est aussi le choix fait par Expectation Maximisation. Cependant, Hierarchical Clustering se trompe et propose un groupe composé seulement du meilleur de tous les élèves, sans que nous comprenions vraiment pourquoi.

La formation A2 (en bleu pointillé) produit des courbes très plates et peu significatives sauf avec K-Means, dont le net point d'inflection à 3 groupes correspond au résultat proposé par X-Means (ce qui est normal dans la mesure où X-Means est basé sur K-Means, même s'il propose en plus un "bon" nombre de groupes).

Pour la formation B, en vert avec des tirets, les quatre méthodes s'accordent pour proposer 3 groupes.

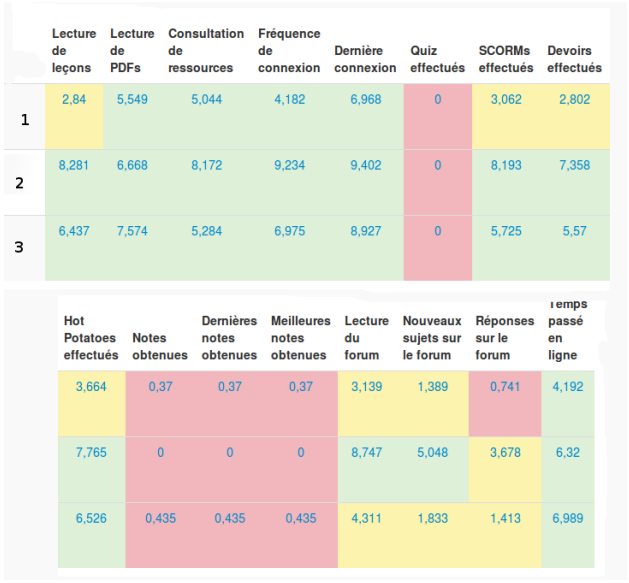
5.2 Signification des groupes

À notre grande surprise, les groupes que nous avons observés pour les trois formations sélectionnées ne montraient guère de distinction plus pertinente que des variations quantitatives entre les élèves plus ou moins actifs, et ce quel que soit le nombre de groupes sélectionné. Par exemple, nous n'avons pas eu de groupe se distinguant d'un autre par son taux d'activité sur le forum comme [9], ni de groupes d'élèves obtenant de bonnes notes malgré une faible activité. Il n'y a eu qu'une formation où certains paramètres permettaient de séparer d'un groupe de bons élèves un sous-groupe où ces élèves faisaient moins d'activités de type devoir. Quand le nombre de groupes requis augmentait à 4 ou 5, plutôt que de voir émerger ce genre de structure, nous obtenions souvent un groupe composé d'un seul élève, souvent le meilleur ou le moins bon de sa formation. Ceci indique que nous atteignions un nombre de groupes trop grand et ne permettant pas bien de représenter les données, et que l'augmenter encore n'aurait pas servi à dégager davantage d'information.

Pour expliquer ce phénomène, nous proposons les raisons suivantes :

- un nombre d'étudiants relativement faible par formation (entre 15 et 56), qui peut présenter moins de variété de comportement ;
- une formation ciblée vers un public peut-être assez homogène en termes d'âge, d'expérience professionnelle, et d'habitudes d'utilisation des TICE ;
- un effet de cercle vicieux pouvant avoir lieu sur le forum faisant que si personne ne se met à l'utiliser, personne d'autre n'aura d'incitation à s'en servir. Nous avons effectivement remarqué parmi nos élèves une tendance à utiliser surtout le forum pour communiquer avec les enseignants plus qu'entre eux.

Cependant, nous avons relevé que la distinction faite entre les élèves pour des nombres de groupes moins élevés était tout à fait pertinente par rapport à notre problématique visant à repérer des élèves en situation d'échec potentiel, surtout dans la mesure où nos résultats semblent indiquer une bonne corrélation entre l'activité globale et les notes obtenues aux activités notées. Utiliser des algorithmes de partitionnement de données est donc bien une solution pertinente à notre problème de suivi des étudiants.



	Lecture de leçons	Lecture de PDFs	Consultation de ressources	Fréquence de connexion	Dernière connexion	Quiz effectués	SCORMs effectués	Devoirs effectués
1	2,84	5,549	5,044	4,182	6,968	0	3,062	2,802
2	8,281	6,668	8,172	9,234	9,402	0	8,193	7,358
3	6,437	7,574	5,284	6,975	8,927	0	5,725	5,57

Hot Potatoes effectués	Notes obtenues	Dernières notes obtenues	Meilleures notes obtenues	Lecture du forum	Nouveaux sujets sur le forum	Réponses sur le forum	Temps passé en ligne
3,664	0,37	0,37	0,37	3,139	1,389	0,741	4,192
7,765	0	0	0	8,747	5,048	3,678	6,32
6,526	0,435	0,435	0,435	4,311	1,833	1,413	6,989

Fig. 4. Un exemple de résultat de partitionnement ne présentant que des différences quantitatives

6 Conclusion et perspectives

Nous présentons un projet visant à utiliser des méthodes d'intelligence artificielle pour répondre au double problème du suivi d'élèves en formation en ligne et de l'amélioration de ces formations. Ce projet est justifié par les difficultés rencontrées actuellement par les responsables de formation pour avoir accès à des statistiques collectives à différents niveaux de granularité représentant l'intégralité des informations disponibles et pas uniquement celles regroupées sur le LMS.

Dans le cadre de ce projet, nous avons été amenés à définir des indicateurs de suivi des élèves qui nous semblent à la fois pertinents à notre projet et suffisamment génériques pour servir à la communauté. Bien que définis à partir de Moodle, nous les pensons généralisables à d'autres LMS.

Nous nous sommes basés sur ces indicateurs pour mener des tests de partitionnement de données sur un jeu restreint de données réelles. Les algorithmes et implémentations choisis pour mener ces tests sont reconnus. Les premiers résultats fournis par ce partitionnement de données sont surprenants, mais nous poussent à approfondir les tests permis par ce nouvel outil.

Nous proposons également une architecture d'application permettant de répondre à notre problématique. Un tel outil est à la fois original et fortement réclamé par les responsables de formation. Nous avons déjà des premiers retours d'usage très positifs de la part des premiers utilisateurs et continuons à améliorer

la plate-forme en conséquence tout en poursuivant l'intégration de nouvelles fonctionnalités.

Cette architecture fait un bon équilibrage entre les différents niveaux de granularité possibles pour l'examen des données : fourniture de données brutes, légèrement traitées pour en faire des statistiques, traitées par des algorithmes classiques puis traitées par des algorithmes innovants dont nous espérons pouvoir montrer à la fois les bons résultats généraux et la bonne adéquation au problème.

La suite du projet sera consacrée à la réalisation de tests plus étendus, à l'implémentation des algorithmes basés sur HTM-CLA, à la mise en place des outils d'amélioration de formation ainsi qu'à la généralisation aux autres LMS.

References

1. Laflaquière, J., Prié, Y.: Des traces modélisées, un nouveau support pédagogique ? In: 4th annual scientific conference - LORNET. (2007)
2. Ngoc, D.P.T., Iksal, S., Choquet, C., Klinger, E.: UTL-CL: A Declarative Calculation Language Proposal for a Learning Tracks Analysis Process. In: Proceedings of the 2009 Ninth IEEE International Conference on Advanced Learning Technologies. ICALT '09, Washington, DC, USA, IEEE Computer Society (2009) 681–685
3. Moodle Trust: Moodle official site (2013) http://moodle.org.
4. Cole, J., Foster, H.: Using Moodle: Teaching with the popular open source course management system. O'Reilly Media, Inc. (2007)
5. Lu, J.: Personalized e-learning material recommender system. In: Proc. of the Int. Conf. on Information Technology for Application. (2004) 374–379
6. López, M., Luna, J., Romero, C., Ventura, S.: Classification via clustering for predicting final marks based on student participation in forums. In: Educational Data Mining Proceedings. (2012)
7. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review (1999)
8. Beaudoin, M.: Learning or lurking? Tracking the "invisible" online student. The Internet and Higher Education (2) (2002) 147–155
9. Taylor, J.: Teaching and Learning Online: The Workers, The Lurkers and The Shirkers. Journal of Chinese Distance Education (9) (2002) 31–37
10. Cobo, G., García, D., Santamaría, E., Morán, J.A., Melenchón, J., Monzo, C.: Modeling students' activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering. In: Educational Data Mining Proceedings. (2011)
11. Romero, C., Ventura, S., García, E.: Data mining in course management systems: Moodle case study and tutorial. Computers and Education (2008) 368–384
12. Calvo-Flores, M.D., Galindo, E.G., Jiménez, M.C.P., Piñeiro, O.P.: Predicting students' marks from Moodle logs using neural network models. Current Developments in Technology-Assisted Education (2006) 1:586–590
13. Numenta: HTM cortical learning algorithms. Technical report, Numenta (2010)
14. Paviotti, G., Rossi, P.G., Zarka, P.: Intelligent Tutoring Systems: An Overview. Pensa Multimedia (2012)
15. Machine Learning Group at the University of Waikato: Weka 3: Data Mining Software in Java (2013) http://www.cs.waikato.ac.nz/ml/weka.
16. Dimitrakopoulou, A.: State of the art on Interaction and Collaboration Analysis (2004) (D26.1.1) EU Sixth Framework programme priority 2, Information society technology, Network of Excellence Kaleidoscope, (contract NoE IST-507838), project ICALTS: Interaction & Collaboration Analysis.