A Learning Analytics Approach to Build Learner Profiles within the Educational Game OMEGA+[*]

Deepak C[1], Maiga Chang[2] and Sabine Graf[2]

[1]Rajalakshmi Institute of Technology, Tamil Nadu, India
[2] Athabasca University, Edmonton, Canada
ocdeepak4@gmail.com; maigac@athabascau.ca; sabineg@athabascau.ca

Abstract. Educational games can act as excellent learning environments, where learners play and learn at the same time. However, typically, once a game has been developed, it is launched and then maybe evaluated for learning effectiveness but details on how learners actually use the game as well as how they play and learn in the game are rarely investigated. In addition, which groups of learners are more attracted or less attracted by the game is seldom looked at. However, such investigations are essential to ensure that the game is used in the way it was intended, that the game is fun and provides learning opportunities at the same time, that learners can benefit the most from the game and to make the game interesting for many different groups of players. In this paper, we introduce a learning analytics approach that builds learner profiles based on learners' characteristics and behaviour in the educational game OMEGA+. The approach is rather generic and can be easily adapted to other educational games. By using the proposed learning analytics approach, clusters of learners are built that provide insights into how learners use the game, how they play and how they learn. In addition, when considering demographic attributes when analysing the clusters, insights can be gained on which groups of learners are more and which groups are less attracted to the game.

Keywords: Game-based learning, Educational games, Learning analytics, Game learning analytics, Clustering, Learner profiling.

1 Introduction

Educational games have high potential in helping students to learn in a fun way. However, similar to online courses, in order to improve such game-based learning environments and ensure that students can benefit most from them, it is important to understand aspects such as how learners use the game, how they behave in it, how much they play in comparison to how much they learn, etc. In addition, understanding which groups of learners/players like the game most and who is not so much attracted

[*] The authors acknowledge the support of Canadian Internet Registration Authority (CIRA)'s Community Investment Program, the National Science and Engineering Research Council of Canada (NSERC) [RGPIN-2020-05837], and Mitacs (Globalink program).

to the game, provides the possibility to expand and tailor the game to those underrepresented player groups to enable them to benefit from the educational game too.

Learning analytics is a fast-emerging research area, which deals with "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs." [1]. Most research in learning analytics looks at online courses as learning environments, several consider social environments (e.g., discussion forums, social network sites, etc.) but only a relatively small number of works conduct learning analytics research in game-based learning environments. (i.e., Alonso-Fernández et al. [2] have conducted a comprehensive systematic literature review on such papers).

In this paper, we propose a learning analytics approach to build learner profiles based on learners' characteristics and behaviours in the educational game OMEGA+ [3][4]. Such profiles can then be used to learn more about how students play and learn in the game as well as how to improve the game design to attract groups of learners that are underrepresented.

According to a systematic literature review on research papers that use learning analytics / data science approaches on game data from educational games, only seven papers exist that focus on building learner profiles based on data from educational games [2]. These papers focus on two directions: First, some research has been conducted on building learner/player profiles to identify certain information from player data (similar to learner modelling). For example, Denden et al. [5] identified personality traits from player behaviour. Another example is the work by Loh and Sheng [6], where authors created a Maximum Similarity Index that represents how (dis)similar the performance of novice players is, compared to expert players within a 'multiple-solution' serious game environment. The other direction includes research on building learner/player profiles based on performance. Examples of such research include the works by Slimani et al. [7], Lazo et al. [8] and Polyak et al. [9], where players are clustered into performance groups.

In this paper, we propose a learning analytics approach that also uses clustering but with the purpose of analysing how learners play, how they learn and how to improve the game to make it more attractive for players/learners who are currently not so attracted to the game. In order to do so, the proposed approach is based on a comprehensive learners/player profile based on multiple learner/player characteristics and behaviours. The proposed approach has been designed for the educational game OMEGA+ but can be easily adapted to other educational games. To verify our approach, a detailed example with simulated data is provided.

This paper is structured as follows: Section 2 provides a brief overview on OMEGA+. Section 3 introduces the proposed learning analytics approach and Section 4 demonstrates the approach through an example. Section 5 then concludes the paper.

2 OMEGA+

OMEGA+ (the former version of the game was called OMEGA) [3][4] is an online educational game that aims at improving four meta-cognitive skills while learners are playing. Those skills are: (1) problem solving, (2) associative reasoning, (3) planning and organization and (4) accuracy and evaluation. In the game, players play matches consisting of a set of three subgames against each other. Overall, there are ten different subgames, each focusing on improving a particular meta-cognitive skill.

In each match, players are scored by how they performed individually through the increase of their meta-cognitive skills and how they performed against their opponent through increase/decrease of their points. For each subgame played, a performance score is calculated that shows how well the player played that subgame. This performance score is then translated into a meta-cognitive skill score of the meta-cognitive skill associated with the respective subgame. To compare players with each other, all performance scores of the subgames within a match are summed up. The winner receives points and the loser loses points, allowing a ranking based on points. The number of won/lost points depends on the overall points the players have before the match.

Besides points and meta-cognitive skill scores, the game entails several other motivational features. Each subgame has 10 difficulty levels where players upgrade to the next level once their average performance over the last 10 times in the subgame is above 70%. Players are also presented with an overall game level, which is the average value of all subgame levels. The game's currency (Ω) is earned for every subgame played depending on how well it is played. If players log in multiple days in a row, they get a bonus to earn more currency per played subgame. Every player is represented by a robot avatar and the earned currency can be used to purchase robot parts to upgrade the player's robot avatar. In addition, a learning analytics dashboard is provided for players to monitor and investigate their game behaviour [10]. Players can unlock 48 badges that are linked to game activities, such as logging in for consecutive days, winning matches, and using the learning analytics dashboard. The game also features a leaderboard with multiple rankings (e.g., by points, metacognitive skill scores, available currency and several other metrics). Players can also send friendship requests to other players. When they play a match, they can then choose to either play against a friend who is currently online or be matched with a random player.

3 Learning Analytics Approach

The proposed approach retrieves relevant data from the game's database and uses a clustering algorithm to classify the data into different groups. Those groups can then be visualized and analysed with respect to their significant characteristics and behaviours to improve our understanding on how players play and learn in OMEGA+ and which groups are more or less attracted to the game. The proposed approach has been implemented in Python using Google's Colaboratory (Colab). The approach consists of four steps, which are explained in the following subsections in more detail.

3.1 Data Retrieval

In this research, we use three different categories of attributes: player details, player possessions, and player activities. Those categories include the following attributes:

Player details.

- GameLevel: The game level presents the average difficulty level the player reached in all subgames.
- Points: Points represent how well a player played matches against other players.
- AgeRange: When creating a player account, players are asked to optionally provide their age range (e.g., 18-24 years, 25-34 years, etc.).
- Gender: Another information that players can provide optionally when creating an account is the gender.
- AllowFriend: This attribute shows whether the player has enabled or disabled friend request. If this option is enabled, other players can send friend request.
- ProblemSolvingSkills: The player's problem-solving metacognitive skills are calculated as a percentage value of his/her performance in the Bypass and Viroid subgames. Only increases in those skills are recorded. All the other metacognitive skills are calculated in the same way.
- AssociativeReasoningSkills: It represents how well the player performs in Associative Reasoning subgames, which are Crossplay, Pattern Hacker and Pirate Hunter.
- PlanningOrganizationSkills: It represents how well the player performs in Planning and Organization subgames, which are CR2k, Evacres and Weekend Barista.
- EvaluateAccuracySkills: It represents how well the player performs in Accuracy and Evaluation subgames, which are Card Swap and Delivery Dash.
- AveragePerformance(1-10): These 10 attributes (one for each subgame) represent the average performance achieved by the player when playing the respective subgame the past 10 times.

Player Possessions.

- Currency: The in-game currency (Ω) is awarded to a player for each played subgame within a match based on their performance and difficulty level.
- RobotParts: This attribute represents the number of robot parts the player has purchased. The player can buy different parts of the robot using in-game currency after fulfilling certain requirements (e.g., earning a certain badge, etc.).
- TotalBadges: This attribute represents the total number of badges a player earned.
- TotalFriends: This attribute represents how many friends a player has in the game.

Player Activities

- TotalMatches: This attribute represents the total number of matches played.
- TotalTime: It represents the total time spent by the player in the game.
- SurveysCompleted: The game contains a few surveys to evaluate the game with respect to its ability to improve meta-cognitive skills of the player, usability and others. This attribute shows whether the player completed any surveys and if so, how many surveys the player completed.
- LeaderboardLog: This attribute represents the number of times the player checks the different leaderboards in the game. More accurately, this attribute counts the

clicks in the leaderboard area that players use to look at different leaderboards or different configurations of the leaderboards.

- AnalyticLog: This attribute represents the number of times the player checks the learning analytics dashboard in the game. The learning analytics dashboard contains (1) line graphs, which show metacognitive skill scores with various filters and visualization options and (2) scatter plots, which show performance scores, again with various filters and visualization options. More accurately, this attribute shows the total number of visualizations created in the learning analytics dashboard.

Code has been implemented that retrieves data regarding the proposed attributes for every player from the game's database.

3.2 Data Preparation

After retrieving the data, it is checked for null values. Some data of attributes in the player possessions and player activities categories may contain null values for some players. Most machine learning algorithms cannot work with missing data [11]. Therefore, any null values were replaced with 0 or mean, depending on the attributes.

In addition, most machine learning algorithms do not perform well when numerical attributes have different scales. Therefore, standardization was used to bring values of different attributes on the same scale. In particular, the StandardScaler transformer feature [12] of the Scikit-Learn library [13] was used to standardize the data. Accordingly, standardization is calculated by subtracting the mean value and then dividing by the standard deviation, so the resulting distribution has unit variance.

3.3 Algorithm

To build learner profiles, the k-means clustering algorithm was used. This algorithm was selected because it is guaranteed to converge, easily adapts to new examples and assigns every player into a cluster [11].

The number of clusters should be specified for the algorithm. To find the optimum number of clusters for the given data, the following steps are performed [11]: first, the model's inertia is calculated for several potential numbers of clusters (i.e., 2 to 9) and plotted on a graph. The inertia of the model is the sum of the squared distance between each instance and its closest centroid [11]. As a result, such graph often contains an inflection point called the elbow after which the inertia decreases much more slowly. Second, another graph is plotted showing the silhouette score of the model for each potential number of clusters. The silhouette score is the mean silhouette coefficient over all the instances [11]. A higher silhouette score is preferred for the optimal number of clusters [11]. Third, by comparing and analysing the elbow value (from the inertia graph) and the silhouette score (from the silhouette score graph), the optimum number of clusters is determined.

Once the optimal number of clusters is determined, the k-means algorithm is executed with that number of clusters and the results of the model, representing which data point belongs to which cluster, is stored.

3.4 Visualization and Analysis

To visualize the high dimensional data, a dimensionality reduction algorithm, namely the Principal Component Analysis (PCA) algorithm [11, 14], is used. The clusters are then visualized in 2D and 3D using python's matplotlib library for a better understanding of the results.

After the results are visualized, each learner in each cluster is analysed and compared with the learners in the same cluster and neighbouring clusters to understand the significant characteristics and behaviour represented in each cluster. In addition, each principal component of the PCA (represented on the axes of the visualizations) is investigated to find out what it represents. Such analysis provides insights into how learners behave in the game, how they play and how they learn. In addition, when looking at demographic attributes such as age range and gender in each cluster, insights can be gained into who is more attracted by the game and who is not.

4 Validation

This section presents a validation of our approach using simulated data from 47 players to demonstrate the different steps in the approach and potential outcomes. The data are not real but were modelled based on the behaviour and activities of beta-testers. As such, the results represent a realistic example to verify our approach, demonstrate how it works and show which kind of insights it can provide.

After the data extraction and preparation, the optimal number of clusters for the k-means algorithm is determined by plotting the model's inertia and silhouette score in a graph for 2 to 9 clusters (see Fig. 1 and Fig. 2). Given that the inertia value decreases slowly after 4 or maybe 5 clusters (Fig. 1) and 4 clusters have a greater silhouette score than 5, the optimum number of clusters for this data is 4. Accordingly, the k-means algorithm is executed using 4 clusters and the results are stored.

Then, the Principal Component Analysis (PCA) dimensionality algorithm is used to transform the high dimensional data into 2D and 3D visualizations (see Fig. 3 and Fig. 4). This is done by identifying the hyperplane that lies closest to the data, and then projecting the data onto it [11].

When analysing our exemplary data, the following can be found: The x-axis may represent some sort of overall activity status in the game, where learners on the lower end are rather passive (i.e., having less matches played, less possessions, less activities, less skills) and learners on the upper end are very active in the game. The y-axis may represent learning effectiveness where learners on the lower end play a lot but improve their skills only little (i.e., high amount of time in the game, high number of matches played, a lot of activities on leaderboards and learning analytics dashboard, but relatively low meta-cognitive skills, relatively low performance in subgames, etc.) while learners on the upper end play little but improve their skills a lot. The z-axis may represent some sort of social status in the game, where learners on the lower end may be less social (i.e., not allow friend requests, have fewer friends, play less matches, complete no or few surveys, etc.) and learners on the higher end or more social.

As such, learners in each cluster may be characterized as follows:

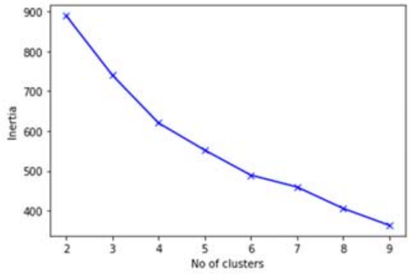

Fig. 1. Inertia Graph

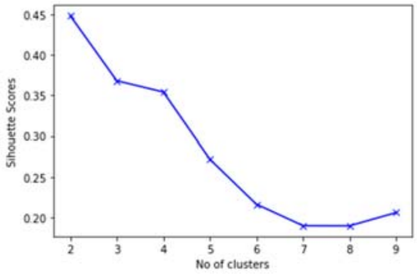

Fig. 2. Silhouette Graph

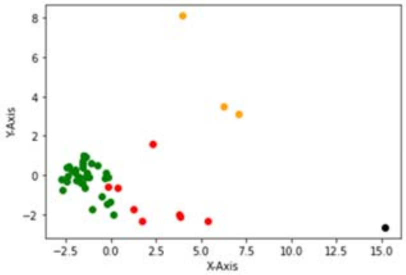

Fig. 3. Visualization of clusters in 2D

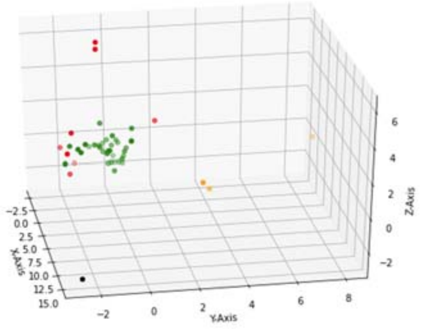

Fig. 4. Visualization of clusters in 3D

- Green Cluster: Learners in this cluster are rather passive given their activity status. They have not played too many matches, do not have a lot of possessions, do not do many activities, and have lower skills. They may be novice players who are not so familiar with the game yet. Their learning effectiveness is low to medium, showing that given the amount of time and activities they do, their skills are somewhat improving. This is again in line with novice players who still need to get familiar with the game. With respect to their social status, they are somewhat social. Again, this is in line with novice players who do not have that many friends yet but are building their social status.
- Red Cluster: Learners in the red cluster are in general more active than learners of the green cluster, however, their learning effectiveness is in general lower. This means that while they seem to spend more time in the game and use different game features, their skills are not improving as much as we would expect. With respect to the social status, some learners are very social while others are not.
- Orange Cluster: Learners in the orange cluster are more active than learners in the red cluster. But in contrast to the red cluster, their learning effectiveness is rather high. This means that while they spend a lot of time in the game and use a lot of its features, they also have high performance in the subgames and improve their skills a lot. With respect to the social status – similar to the red cluster – some learners are quite social while others are not.

- Black Cluster: Only one learner was assigned to this cluster. This learner seems to be extremely active, but his/her learning effectiveness is rather low. This means that although the learner spends a lot of time in the game and uses a lot of features, he/she does not improve his/her skills as it would be expected. This could be because he/she might be more distracted by some of the features (e.g., spending hours on looking through different leaderboards).

While those descriptions of axes and clusters are just exemplary, they demonstrate well how powerful this approach can be in finding out more about how learners use and play in the game and how/whether they benefit from the game. In addition, demographic attributes such as age range and gender can be used to further analyse the clusters (if those attributes are not already dominant in the principal components).

For example, we may see in the data that the distribution of male and female players is similar in the green cluster, where we have mainly players who just started to play and/or are not very active in the game. However, when looking at the other clusters, where we have players who are more active, we see that the percentage of female learners compared to male learners is significantly lower than it is in the green cluster. This may show that female players are not as active in the game and do not benefit much from the game due to their low activity status. Such findings can then be used to improve the game design to attract those learner groups (i.e., female learners) and add features that may make the game more interesting for them.

5 Conclusions

This paper presents a learning analytics approach to build learner profiles in the educational game OMEGA+. The profiles are created through cluster analysis and consider a variety of features related to learners' characteristics, possessions in the game and their activities in the game. The approach has been validated with simulated data from 47 players to demonstrate the insights and benefits this approach can provide.

Most related works focus either on identifying new information (e.g., personality traits, new performance metrics, etc.) from behaviour in an educational game [e.g., 5, 6] or on clustering based on performance in an educational game [e.g., 7-9]. However, the clusters in this approach are built by considering not only performance but a variety of other learner characteristics, their possessions in the game as well as their activities in the game. By considering such a diverse set of attributes when building the clusters/groups of learners, insights into how learners use the game, how they play and how they learn can be gained. In addition, by considering demographic attributes, investigations can be conducted into the attractiveness of the game for different groups of learners. Such insights can be used to improve the game design, on one hand, to ensure that it is used the way it was intended and really provides learners with learning opportunities that are fun for them and, on the other hand, to broaden the reach of the game and make it attractive for more diverse groups of learners.

Future work will deal with using our approach on real data to learn more about the effectiveness and reach of OMEGA+. In addition, future work will deal with adapting our approach to other educational games and using it with real data for those games.

References

1. Siemens, G., Gašević, D.: Special Issue on Learning and Knowledge Analytics, Educational Technology & Society, 15(3), 1–163 (2012).
2. Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., Fernández-Manjón, B.: Applications of data science to game learning analytics data: A systematic literature review, Computers & Education, 141, 103612 (November 2019).
3. Chang, M., Graf, S., Corbett, P., Seaton, J., McQuoid, S., Ross, T.: OMEGA: A Multiplayer Online Game for Improving User's Meta-Cognitive Skills. In Proceedings of the IEEE International Conference on Technology for Education (T4E 2019), pp. 178-185. IEEE Computer Society, Goa, India (2019).
4. OMEGA+, https://omega.athabascau.ca, last accessed April 12, 2022.
5. Denden, M., Tlili, A., Essalmi, F., Jemni, M.: Implicit modeling of learners' personalities in a game-based learning environment using their gaming behaviors, Smart Learning Environments, 5, 29 (November 2018).
6. Loh, C. S., Sheng, Y.: Maximum Similarity Index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games, Computers in Human Behavior, 39, 322-330 (October 2014).
7. Slimani, A., Elouaai, F., Elaachak, L., Bakkali Yedri, O., Bouhorma, M., Sbert, M.: Learning Analytics Through Serious Games: Data Mining Algorithms for Performance Measurement and Improvement Purposes, International Journal of Emerging Technologies in Learning, 13(1), 46-64 (2018).
8. Lazo, P. P. L., Anareta, C. L. Q., Duremdes, J. B. T., Red, E. R.: Classification of public elementary students' game play patterns in a digital game-based learning system with pedagogical agent. In Proceedings of the 6th International Conference on Information and Education Technology (ICIET 2018), pp. 75–80. ACM, New York, NY, USA (2018).
9. Polyak, S. T., von Davier, A. A., Peterschmidt, K.: Computational Psychometrics for the Measurement of Collaborative Problem Solving Skills, Frontiers in Psychology, 8 (2017).
10. Seaton, J., Chang, M., Graf, S.: Integrating a Learning Analytics Dashboard in an Online Educational Game. In Chang, M., Tlili, A. (eds.) Data analytics approaches in educational games and gamification systems, pp. 127-138. Springer, New York (2019).
11. Géron, A.: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow - Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd edn. O'Reilly Media, Inc., Canada (2019).
12. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Muller, A. C., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project, arXiv preprint arXiv:1309.0238 (2013).
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011).
14. Jolliffe I. T., Cadima, J.: Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society A - Mathematical, Physical and Engineering Sciences, 374(2065) (April 2016).