

REVIEW OF LEARNING ANALYTICS AND EDUCATIONAL DATA MINING APPLICATIONS

Eleni Mangina, Georgia Psyrra

School of Computer Science – University College Dublin (IRELAND)

Abstract

During the past decade the educational data research is rapidly growing. The use of technology in education has created the need to store and manage large amounts of data that come from various sources and have different formats. Educational data can be used to benefit educational systems and the science of learning. The project “Augmented Reality Interactive Educational System” (ARETE) funded by EU Programme Horizon 2020 aims to support interactive technologies for the provision of Augmented Reality (AR) content through an open source learning management system and authoring toolkit for the broader community of users. The utilisation of educational data is vital for the efficient data management and the two relevant areas of focus in review that focus on the use of educational data to support education are the Educational Data Mining (EDM) and Learning Analytics (LA). Several studies have been published recently focusing on applications using educational data, revealing that educational data analytics is an evolving science, where researchers have explored the various use cases of applying data mining and analytics techniques on the educational domain. However, there is still a need for exploring the main objectives of applying EDM and LA techniques and defining the specific problems in the educational domain they try to resolve. The aim of this analysis is to identify studies’ objective trends that recent applications are trying to achieve and to identify potential research gaps. The possible correlation between the use of particular types of techniques used by the EDM/LA applications in relation to the goals they are trying to achieve is also being presented. This paper presents the review of EDM and LA empirical studies that have been published between 2016 to 2020. To gain insight into the trend direction of the different projects, the publications are clustered based on the methods applied and the purposes those studies tried to accomplish. Studies that applied more than one technique were assigned to the method groups more than once. This paper will provide an association table of EDM/LA techniques and the objectives for which they have been used, and will serve as a model for other researchers in order to choose the method for their own specific goals. Finally, the goals that recent EDM and LA applications are approaching will be presented, which can be a source of inspiration for further research questions, by providing information on areas of educational goals that remain unexplored or have not received much attention so far.

Keywords: Educational Data Mining, Learning Analytics.

1 INTRODUCTION

EDM and LA are increasingly popular fields of research as they target to improve every aspect of the education process. According to Romero and Ventura (2020) [1], in the bibliography the terminology is referred within a wide range of related terms such as Academic Analytics, Institutional Analytics, Teaching Analytics, Data-Driven Education, Data-Driven Decision-Making in Education, Big Data in Education, and Educational Data Science. Both educational analytics fields are linked with computer, statistic and education related sub areas. A presentation of their relationship with these various science fields is illustrated in the figure below.

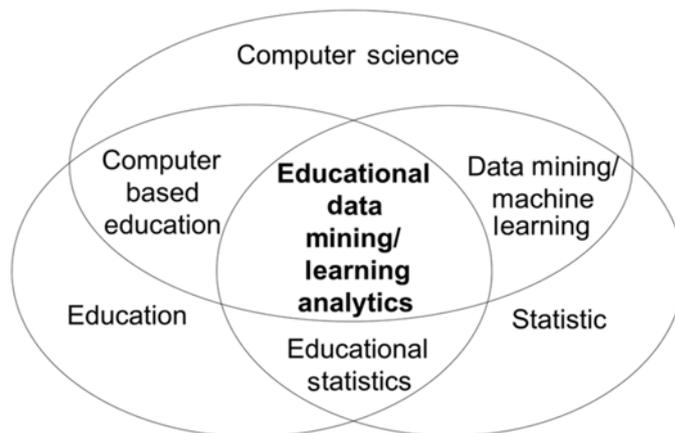


Figure 1: Areas related to Educational Data Mining/Learning Analytics

Source: (Romero & Ventura, 2020) [1]

EDM and LA are research fields which include the processes of collection, analysis and interpretation of educational data [2]. From a general scope EDM concerns more about enhancing the techniques and methodologies used to analyse educational data, while LA focuses more on their applications [3].

Siemens & Baker (2012) [4] argued that the origin difference of EDM and LA is that the EDM is linked with educational software and student modelling, while LA is more related to the semantic web, “intelligent curriculum”, outcome prediction and systemic interventions. Nevertheless, they conclude that these differences seem to be less obvious as both fields continue to evolve.

Bienskowski and their co-authors (2014) [2] in their work distinguish the EDM and LA by the education layer that each focuses on. They demonstrated that EDM has the aim to discover micro-concept patterns by developing new tools while LA works in a larger spectrum by applying tools and techniques. However, they also argued that both fields have similar objectives such as to focus on discovering patterns and using them in order to create accurate predictions.

For the needs of the ARETE H2020 project [5] [6], we have reviewed the methods and techniques that EDM and LA utilised for consideration of the future development of the ARETE platform.

2 EDM AND LA COMMON METHODS

EDM and LA are both fields of educational data analytics and share common methods and techniques and achieve similar objectives. This section presents a classification of the studies collected based on the methods applied. Most of the studies applied multiple techniques, hence those studies were assigned to more than one method group. Calvet Liñán & Juan Pérez (2015) [3], in their work “Educational Data Mining and Learning Analytics: differences, similarities, and time evolution” demonstrated the following LA and EDM common methods: Prediction, Clustering, Relationship mining, Distillation of data for human judgment, Discovery with models, Outlier detection, Social network analysis, Process mining, Text mining, Knowledge tracking and Non-negative matrix factorization. 33 open source studies were found and analysed, and based on the methods they use to achieve their goals the main methods utilised were: Relationship mining [7], Prediction [8], Process Mining [9], Clustering [10], Social Network Analysis (SNA) [9] [11] and Text mining [9]. The count of articles per method used shows that less studies applied text mining and SNA techniques.

Under the science of Educational Data Analytics, EDM and LA are similar fields and approach the analysis of educational data in order to improve educational experiences. Andy Peterson (2018) [12] distinguished the applications and data their approaches wish to implement. He claimed that LA target to capture students’ behaviour, performance and interactions with the learning environment while EDM tries to explore patterns that aim to represent student achievement and performance in the academic environment.

Papamitsiou and Economides (2014) [13], in their study on the use of learning analytics and data mining in education, explored the variety of the main objectives of the major research conducted between 2008 and 2013. The common goals that they found to be realized by educational data analysis studies in that

period are: the modelling of students' behaviour, the prediction of their academic performance, the increase of (self-) reflection and (self-) awareness, the prediction of dropout and retention, the improving of feedback and evaluation services and the recommending resources.

Additionally, Calvet Liñán and Juan Pérez (2015) [3] demonstrated some common EDM and LA key applications some of which can be summarized based on the most popular analytical methods we found as follows:

- Prediction. Prediction and detection of students' behaviour.
- Clustering. Grouping of similar materials or students based on their learning and interaction patterns.
- Relationship mining. Identifying relationships in learner behaviour patterns and diagnosing students' difficulties.
- SNA. Interpretation of the structure and relations in collaborative activities and interactions with communication tools.
- Process Mining. Reflecting student behaviour in terms of its examination traces, consisting of a sequence of course, grade and timestamp.

3 RESULTS

In order to investigate the main objectives of the recent educational data analysis studies in this literature, covering the period 2016-2020, as well as to relate them to the method used, the studies were classified based on the categories of techniques used in relation to their objectives. This classification (see Table 1) seeks to offer the reader a general overview of the techniques used in particular domains.

Table 1. EDM/LA Objectives according to the techniques used.

<i>Technique</i>	<i>Purposes / Domain</i>	<i>Count</i>
Regression	Prediction of students' grade value. Detection of performance predictors.	3
Probability estimation	Learning Objects Recommendation system.	1
Classification	Students' upcoming courses performance prediction. Estimation of dropping out.	3
	Learning materials classification in order to support learning personalization	1
	Students performance prediction. Drop out detection.	3
	Knowledge change prediction. Whether or not to move to the next activity.	2
	Classify teaching assistance workforce under performance ability categories.	1
	Academic integrity detection / Cheating detection	1
	Students' behaviour and learning strategy labeling.	3
Model discovery	Change of knowledge estimation through progress tracing.	3
	Students performance prediction. Drop out detection.	4
	Students' behaviour and learning strategy detection.	2
Hard clustering	Performance predictors estimation for detecting students at risk.	2
	Association of teachers and students.	1
	Detect students' behaviour and learning paths.	2
Association rules	Performance patterns at the students' knowledge level.	1
	Relationships of students' mistakes in exercises.	1
	Detection of learning patterns to create high-level student profiles.	1
	Context aware and rationalization.	1
	Determining performance relationships of students who often fail compared to those who almost never fail.	1
Correlation mining	Estimation of students' performance predictor factors.	2
Distillation of data for human judgment	Providing feedback of students behaviour and attendance.	4

Sequential pattern mining	Students' behaviour and learning strategy detection.	2
	Change of knowledge estimation through progress tracing.	2
	Grade prediction using temporal behaviour data.	1
Knowledge tracing	Students' behaviour and learning strategy detection.	1
	Change of knowledge estimation through progress	2
Causal mining	Identification of academic performance predictors.	4
Social network analysis	Students' performance prediction based on social interactions or co-enrolment networks.	2
	Interpretation of collaborative activities structures.	1
Text classification	Academic integrity detection / Cheating detection	1

Based on the frequency of studies found under the same pair of objectives and used method and additionally having a limit of at least 3 studies to be included in that group, we observed the following five pairs of EDM/LA techniques and studies purposes:

- Prediction/Regression analysis – Prediction of students' grade value. Detection of performance predictors. Romero and Ventura (2013) [9] argued that the prediction method used by EDM for students' performance estimation and detection of students' behaviour applications. They additionally presented 3 types of prediction including the classification which was used to predict a categorical value, the regression which focused on approaching continuous values and the density estimation which was used to predict a probability density function.
- Prediction/Classification – Students performance prediction. Estimation of dropping out. Students' behaviour and learning strategy labeling. More than one objectives category was assigned to the classification prediction technique. With a general perspective we can deduce that those goals are related with students' performance classification and estimation of dropping out, students' behaviour and learning strategy labelling, learning materials classification, knowledge change prediction, academic integrity detection and classification of teaching assistance workforce under performance ability categories.
- Process mining/Model discovery – Students' behaviour and learning strategy detection. Students performance prediction. Drop out detection. According to Cairn et al. (2015) [14] educational process mining is a rising field of EDM that targets on developing methods to be more familiar with students' behaviour and the factors that affect their performance. They also argued that process mining deals with the revelation, analysis and provision of a visual representation of integrated educational processes. Romero and Ventura (2013) [9] noted that use of the process mining technique in an educational context can achieve representative visualizations of students' habits and behaviour in terms of examination traces consisting of a sequence of course, academic performance, and timestamp for each student.
- Relationship mining/Distillation of data for human judgment– Providing feedback of students behaviour and attendance. A relationship mining technique used to represent data in intelligible ways using summarization, visualization, and interactive interfaces [1]. The studies assigned to this technique group focused on providing feedback of students behaviour and attendance.
- Relationship mining/Causal mining – Identification of academic performance predictors. Aims to find causal structures among variables, meaning that one variable causing another variable [15]. The studies found in our literature collection to apply that method focused on identifying the factors that affect students' academic performance.

3.1 EDM/LA studies' Goals and Objectives

The table presented above (see Table 1) reveals that different EDM and LA techniques can be used for similar purposes. Thus, in order to find the popular goals of recent EDM and LA studies, we classified the studies regardless of the used method. Using this approach, we intensified 15 general goal categories that recent studies have focused on.

It is concluded that among the goals of EDM and LA applications, 46.46% are related with students performance prediction and its predictor factors as well as with students behaviour and learning strategy

detection; 33.30% are related with student profiling and detection of students at risk analyses and 26.67% are related with students attendance and behaviour feedback. Additionally, fewer studies' objectives were focused on the domains of the prediction of students' grade value (20.0%); change of knowledge estimation (13.3%) and learning materials recommendation (13.3%). Even less (6.6%) had objectives related with classification of teachers' performance and association with students, academic integrity detection, students' performance patterns and context awareness for personalized learning.

4 CONCLUSIONS

The development in the emerging fields of educational data mining and learning analytics as well as the variety of methods and achievement goals on which their applications focus on can be seen from the availability of literature from 2016 until today. From the 33 articles selected for review 20 were based on relationship mining methods and focus on objectives such as to provide feedback of students' behaviour and attendance and to identify academic performance predictors. The next most common method the studies used was prediction analysis (18) which included objectives such as prediction of students' grade value, detection of predictor factors and dropping out estimation. Fewer studies were based on process mining (9) and clustering (5) and targeted on objectives such as students' behaviour and learning strategy detection and performance predictors estimation for detecting students at risk. Even less on social network analytics (3) and text mining (1). Finally, we observed that it is not common for an EDM/LA study to use one particular method. The number of studies which applied a combination of methods is above average, thus the results above are not referred to an exclusive use of one method. For the needs of the ARETE Authoring Toolkit platform and the collection of educational data, further development will utilise Augmented Reality metadata (Secretan et al, 2019) [16] and Learning Objects' metadata and methods identified within this review will be considered for analysis of the metadata provided from the Learning Management System [17].

ACKNOWLEDGEMENTS

The publication has been supported by European Union's Horizon 2020 research and innovation program under grant agreement No 856533, project ARETE.

REFERENCES

- [1] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020. doi: <https://doi.org/10.1002/widm.1355>
- [2] Bienkowski M, Feng M, Means B. Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief. Office of Educational Technology, US Department of Education, 2012.
- [3] L. Calvet Liñán and Á. A. Juan Pérez, "Educational Data Mining and Learning Analytics: differences, similarities, and time evolution," *International Journal of Educational Technology in Higher Education*, vol. 12, no. 3, pp. 98-112, 2015. Retrieved from <https://www.raco.cat/index.php/RUSC/article/view/v12n3-calvet-juan/394311>
- [4] G. Siemens and R. S. J. D. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in Proceedings of the 2nd international conference on learning analytics and knowledge, LAK 2012. pp. 252-254, 2012. doi: 10.1145/2330601.2330661.
- [5] S. Masneri et al., "Work-in-progress-ARETE-An Interactive Educational System using Augmented Reality," in 6th International Conference of the Immersive Learning Research Network, iLRN 2020. pp. 283–286, 2020. doi: 10.23919/iLRN47897.2020.9155186.
- [6] Zenodo, H2020 ARETE Project, Accessed 18 April, 2021. Retrieved from <https://zenodo.org/communities/augmented/?page=1&size=20>
- [7] M. Anoopkumar and A. M. J. Md Zubair Rahman, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration," in International Conference on Data Mining and Advanced Computing, SAPIENCE 2016. pp. 122-133, 2016. doi: 10.1109/SAPIENCE.2016.7684113.

- [8] M. Kumar, A. J. Singh, and D. Handa, "Literature Survey on Educational Dropout Prediction," *International Journal of Education and Management Engineering*, vol. 7, no. 2, pp. 8-19, 2017. doi: 10.5815/ijeme.2017.02.02.
- [9] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12-27, 2013. doi: 10.1002/widm.1075.
- [10] S. Kaushik, An Introduction to clustering and different methods of clustering, *Analytics Vidhya*, 3 November, 2016. Retrieved from <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>
- [11] D. Z. Grunspan, B. L. Wiggins, and S. M. Goodreau, "Understanding classrooms through social network analysis: A primer for social network analysis in education research," *CBE-Life Sciences Education*, vol. 13, no. 2, pp. 167-78, 2014. doi: 10.1187/cbe.13-08-0162.
- [12] A. Peterson, "Chapter 13 - Big Data in Education: New Efficiencies for Recruitment, Learning, and Retention of Students and Donors," in *Handbook of Statistical Analysis and Data Mining Applications*, pp. 259-277, Charlotte: Elsevier, 2018. Retrieved from <https://doi.org/10.1016/B978-0-12-416632-5.00013-X>
- [13] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systemic literature review of empirical evidence," *Educational Technology & Society*, vol. 17 no. 4, pp. 49-64, 2014. Retrieved from <https://www.jstor.org/stable/jeductechsoci.17.4.49>
- [14] A. H. Cairns, B. Gueni, M. Fhima, A. Cairns, S. David, and N. Khelfa, "Process Mining in the Education Domain," *International Journal on Advances in Intelligent Systems*, vol. 8 no. 1, pp. 219-32, 2015. Retrieved from [72.52.166.99/articles/intsys_v8_n12_2015_18.pdf](https://www.elsevier.com/locate/ijais)
- [15] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable techniques for mining causal structures," *Data Mining and Knowledge Discovery*, vol. 4, no. 2–, pp. 163–192, 2000. doi: 10.1023/A:1009891813863.
- [16] J. Secretan, F. Wild, and W. Guest, "Learning Analytics in Augmented Reality : Blueprint for an AR / xAPI Framework," in *IEEE International Conference on Engineering, Technology and Education, TALE 2019*. pp. 1-6, 2019. doi: 10.1109/TALE48000.2019.9225931.
- [17] ARETE, Moodle Digital Repository, Accessed 18 April, 2021. Retrieved from: <https://arete.ucd.ie/>