A Brief Review of Educational Data Mining to Improve Intelligent Learning Environments

Mayra Mendoza, Yasmín Hernández, Javier Ortiz, Alicia Martínez, Hugo Estrada

Tecnológico Nacional de México, Cuernavaca, Morelos, Mexico

{m21ce017, yasmin.hp, javier.oh, alicia.mr,hugo.ee}@cenidet.tecnm.mx

Abstract. Learning environments and educational platforms have become ubiquitous in current education. These systems produce an increasing volume of data about different aspects of education. The analysis of this data yields useful knowledge to understand the interrelations and individual states of actors in education, such as students, pedagogical actions, tutorial decisions, learning strategies and learning outcomes. These insights can be used in many ways to improve learning and education, for example, improving intelligent tutoring systems and intelligent learning environments. We are developing two intelligent tutoring systems and we are analyzing data from educational contexts to model the different components of the intelligent tutors through applying several educational data mining techniques. The initial mining approach is presented along with a brief review of literature on educational data mining.

Keywords: educational data mining, intelligent tutoring systems, intelligent learning environments, student modeling.

1 Introduction

Traditionally, education has been one of the favorite fields to prove computational theories with the aim of supporting the teaching and learning processes; for example, the artificial intelligence has produced several applications to improve the learning process through adaptive teaching and tutoring. As a result, there are several educative platforms such as intelligent tutor systems (ITS), learning management systems (LMS), e-learning systems, serious and educational games, and massive open online courses (MOOC), in addition to administrative computer-based systems. These educative

platforms have produced a growing volume of data about the interaction of students with learning environments, student preferences, student states, tutorial decisions, tutorial situations, pedagogical strategies, and their impact in learning. The educational data can help to know the students, to understand different aspects of the interaction of the students with the systems, and to comprehend the learning process itself, and therefore to improve the systems and processes.

The Educational Data Mining (EDM) is an emerging discipline, interested in the development of methods to explore the exceptional data that comes from educational environments, and concerned in the use of these methods to understand students and the environments in which they learn [10]. EDM is based on statistical methods and machine learning algorithms.

There is extensive research to understand educative processes. On the one hand, researching is interested in knowing which are the characteristics of students with more impact in learning; for example, emotions, motivation, self-efficacy, among other characteristics, or to know the context or circumstances where the tutorial/teaching actions are successful, and therefore to improve the learning environments. On the other hand, research is also interested in obtaining useful knowledge for teachers, pedagogists, education managers, researchers in educational psychology and learning sciences, and other stakeholders to improve their functions. For example, the prediction of drop-out is a very investigated problem.

We are developing two intelligent learning environments, the first is intended to teach math to kids from elementary school, and the second one is being designed to support undergraduate and graduate students to learn mathematical logic. We want to know which aspects of the learning should be modeled to improve. Therefore, we are analyzing several educational datasets with EDM techniques. In a first stage, we are analyzing public datasets, but also, we are planning to gather our own data. Mainly, we are interested in understanding the impact of psychological constructs in learning, such as motivation, self-efficacy, and self-regulated learning.

We present a brief review of relevant work on EDM, and we depict a proposal to apply educational data mining to understand important factors in learning. The paper is organized as follows: section 2 presents a description of the EDM field; section 3 presents a brief review of literature on educational data mining; section 4 depicts our approach for applying data mining techniques on educative contexts. Finally, section 5 outlines our conclusions and future work.

2 Educational Data Mining

Nowadays, plenty of data is generated, since almost every aspect of our daily life can be tore apart in pieces of information. The imperative necessity of solving problems led to the analysis of the growing data and gave birth to data mining. Data mining is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage [11].

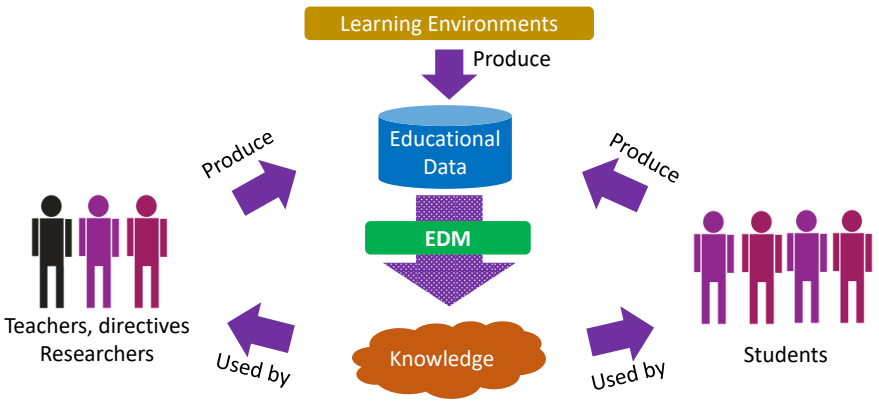

Fig. 1. Educational Data Mining knowledge discovery cycle [10].

Machine learning is the technical basis in data mining. Machine learning is concerned with the ability of a system to acquire and integrate new knowledge through observations of users and with improving and extend itself by learning rather than by being programmed with knowledge [11]. These techniques organize existing knowledge and acquire new knowledge by intelligently recording and reasoning about data. For example, observations of the previous behavior of students will be used to provide training examples that will form a model designed to predict future behavior.

Data from educational contexts can be analyzed to obtain knowledge about learning and students, and to have a better, smarter, more interactive, engaging, and effective education. Educational Data Mining (EDM) is an emerging research field concerned with the application of data mining, machine learning and statistics to data generated by educational settings (schools, universities, Intelligent Tutoring Systems, Learning Management Systems, and MOOCs). EDM seeks to develop and improve methods for exploring this data to discover insights about how people learn. EDM still has many pending issues; but it has the potential to support the development of other fields related to education. This requires advances in artificial intelligence and machine learning, human intelligence understanding and learning theories [7].

As in data mining, in EDM several computing paradigms and algorithms converge, such as decision trees, artificial neural networks, machine learning, Bayesian learning, logic programming, statistical algorithms, among others. However, traditional mining algorithms needs to consider the characteristics of the educational context to support instructional design and pedagogical decisions [10].

Educational data has meanings with multiple levels of hierarchy, which need to be determined by means of the properties of the data itself. Time, sequence, and context play an important role in the study of educational data. EDM supports the development of research on many problems in education, since it not only allows to see the unique learning trajectories of individuals, but it also allows to build increasingly complex and sophisticated learning models [4].

The knowledge uncovered by EDM algorithms can be used not only to help teachers manage their classes, understand learning processes of their students, and reflect it in

their own teaching methods, but also to support reflections of the student about the situation and give feedback to them [8]. Although one might think that there are only these two stakeholders in EDM, there are other groups of users, who see EDM from different points of view, according to their own objectives [10]. For example, education researchers, universities, course developers, training companies, school supervisors, school administrators, could also benefit from the knowledge generated by EDM [7]. Fig. 1 shows the interrelationships of educational environments, stakeholders and the EDM process.

3 Review of Educational Data Mining Research

The improvement of computing has admitted to store and process huge data which some years ago was impossible. As a result, educational technologies have been instrumented to collect large amounts of data which in turn is analyzed to understand the several aspects and interrelations of the educative processes.

Many researchers are interested in the development of early detection of struggling students. For example, Hung and colleagues [6] propose a novel predictive modeling method to address the research gaps in existing performance prediction research., their focus is on:

i) the lack of existing research focused on performance prediction rather than identifying key performance factors
ii) the lack of common predictors identified for both K-12 and higher education environments
iii) the misplaced focus on absolute engagement levels rather than relative engagement levels.

The predictive modeling technique was applied in two datasets, one from higher education and the other from a K-12 online school with 13,368 students in more than 300 courses. Some experiments were conducted. First, because a student's engagement level has been identified as a key predictor in predicting performance, the input variables will be converted from absolute values to relative values. A K-fold model for training and validation will be applied to compare accuracy and recall rates between absolute and relative models.

Authors assume the prediction accuracy can be improved by the relative transformation. Secondly, because prediction errors are inevitable in the process of predictive modeling, they propose a two-stage approach by constructing three related predictive models (the successful model, the at-risk model, and the coordination model).

Authors assume the dual-stage, three-model approach can further improve the model's accuracy and capture more at-risk students during the semester. Finally, because ensemble and deep learning models will be adopted as the major algorithms, the surrogate model approach will be applied to reveal the key at-risk predictors. The results will assist in identifying common at-risk predictors across K-12 and higher education learning environments [6].

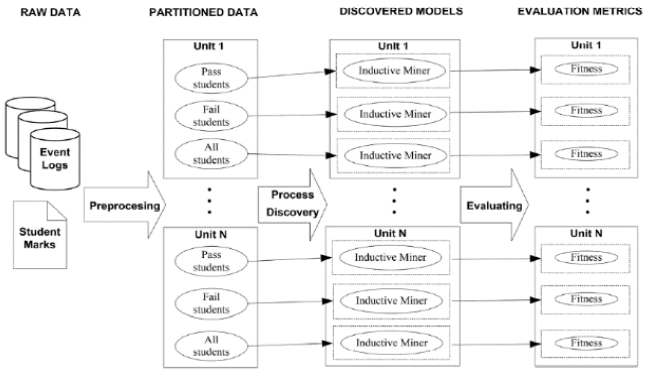

Fig. 2. Educational Processes Mining process from raw data to algorithm interpretation [2].

The results showed the newly suggested approach had higher overall accuracy and sensitivity rates than the traditional approach. In addition, two generalizable predictors were identified from instruction-intensive and discussion-intensive courses. Authors suggest that for future research, researchers should consider applying this dual-stage approach to other predictive modeling tasks. Combining online discussion content and online behaviors, reflected by student efforts in online courses, might further improve the analytical results reported [6].

On the other hand, the content assessment through educational data mining has received attention and it has also been broadly improved in e-learning scenarios in recent decades. Cerezo and colleagues [2] state that the e-Learning process can give rise to a spatial and temporal gap that poses interesting challenges for assessment of not only content, but also acquisition of core skills such as self-regulated learning. Their research is focused in to discover self-regulated learning processes of students during an e-Learning course by using Process Mining Techniques.

They applied a new algorithm in the educational domain called Inductive Miner over the interaction traces from 101 university students in a course given over one semester on the Moodle 2.0 platform. Data was extracted from the platform's event logs with 21,629 traces to discover self-regulation models of students that contribute to improving the instructional process.

The Inductive Miner algorithm discovered optimal models in terms of fitness for both Pass and Fail students in this dataset, as well as models at a certain level of granularity that can be interpreted in educational terms, which are the most important achievement in model discovery.

Authors conclude that although students who passed did not follow suggestions of the instructors exactly, they did follow the logic of a successful self-regulated learning process as opposed to their failing classmates. They state that the Process Mining models allows them to examine which specific actions were performed by the students. They found interesting a high presence of actions related to forum-supported collaborative learning in the Pass group and an absence of those in the Fail group [2]. This process is depicted in Fig. 2.

Another research is interested is measure several psychological constructs. For example, Li and colleagues are concerned in developing models to measure self-regulated behavior and identify significant behavioral indicators in computer-assisted language learning courses. In their models, the behavioral measures were based on log data from 2454 freshman university students from Art and Science departments for a year. These measures reflected the degree of self-regulation, including anti-procrastination, irregularity of studying intervals, and pacing.

Authors apply clustering analysis to identify typical patterns of learning pace, and hierarchical regression analysis was performed to examine significant behavioral indicators in the online course. The results of learning pace clustering analysis revealed that the final course point average in different clusters increased with the number of completed quizzes, and students who had procrastination behavior were more likely to end up with lower final course points. Furthermore, the number of completed quizzes and studying intervals irregularity were strong predictors of course performance in the regression model. This clearly indicated the importance of self-regulation skills, in particular the completion of assigned tasks and regular learning.

In the context of intelligent tutoring systems research, the student model has received more attention that the other components since it enables the ITS to respond to the needs of the students. The student modeling has been improved due to educational data mining. These authors build a student model based on the data log of a virtual reality training system that has been used for several years to train electricians. They compared the results of this data-driven student model with a student model built by an expert. For the knowledge representation, authors rely on Bayesian networks to build the student models.

Bayesian networks have been used in ITS to model student knowledge, predict student behavior and make tutoring decisions due to their strong mechanisms for managing the involved uncertainty. The model relies on Bayesian networks to probabilistically relate behavior and actions of the students with their current knowledge. The tree augmented naive Bayes algorithm and the GeNIe software package were used to learn the Bayesian model from the data from the system for electrical training. They conducted an initial evaluation comparing the data-driven student model with a student model built with expert knowledge. And both models obtain good results in predicting the student state [5].

Another plentiful source of educational data are forums, chats, social networks, assessments, essays, among others, which produce a massive volume of data, especially in text format. According to Ferreira and colleagues [3], documents pose exciting challenges on how to mine text data to find useful knowledge for educational stakeholders. These authors conducted a systematic overview of the current state of the Educational Text Mining field. Their final goal is to answer three main research questions: Which are the text mining techniques most used in educational environments? Which are the most used educational resources? And which are the main applications or educational goals? Authors state although there is much research published in educational text mining, it also has gaps to be filled in, and there are some hot and new topics to develop. More specifically, they propose the next most interesting future research lines:

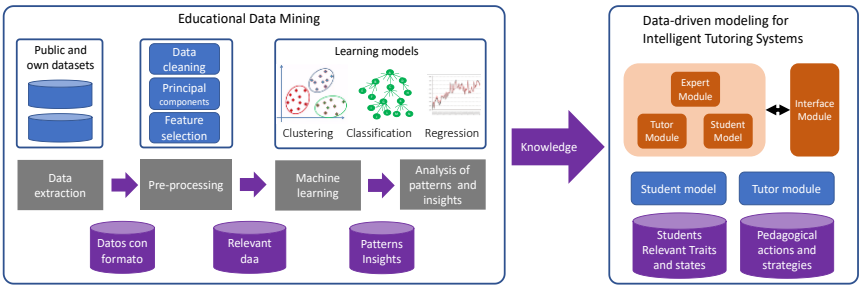

Fig. 3. Data-driven approach to intelligent tutoring systems modeling.

i) analyzing online discussion collaboration,
ii) writing analytics,
iii) natural language generation.

4 Applying EDM to Improve Learning Environments

Educational data mining plays an important role in understanding the different aspects of the learning process and in building and improving intelligent learning environments. These techniques organize existing knowledge and acquire new knowledge by intelligently recording and reasoning about data. For example, observations of the previous behavior of students will be used to provide training examples that will form a model designed to predict future behavior of students [12].

We are developing two intelligent tutoring systems (ITS). The first one aims to teach math to kids of elementary school. The second one targets to tech mathematical logic to undergraduate and graduate students. We want to design the components of the ITS considering the knowledge generated by data coming from educational contexts. We are using a standard approach for data mining, which is shown in Fig. 3.

We are following two strategies. The first one consists of analyzing public datasets and the second one is interested in gathering our own data. In this moment, we are analyzing several public datasets at repositories as DataShop, which is a big repository of learning interaction data. Such data can be used to help advance our understanding of student learning and learning process itself [7].

We are applying techniques for feature selection and principal components analysis, to detect those relevant attributes for learning and to eliminate those attributes which do not provide information to the model. The datasets which we are working on are diverse and heterogenous, have many attributes and they were gathered with different purposes. Therefore, we are training the predictive models using algorithms for clustering, classification, and regression.

With the knowledge generated by the EDM process, we are designing the intelligent tutoring systems. In this first stage, we are working on the student model, and in the tutor module, therefore, we will have data-driven student models. The student model is an important component of the ITS because it allows the ITS to provide adaptive

instruction to students; and the tutor module make decisions about the tutorial actions to be presented to the students based on the student model.

As first steps, we need to identify the relevant attribute in students which have positive impact in learning. In our previous research, we have identified emotions, personality, and goals as important players in motivation and learning. But now, we want to explore other relationships. We are trying to model self-efficacy and self-regulated learning; therefore, we are working in identifying the indicators of self-efficacy and self-regulated learning to include them in the student model of the ITS.

Self-efficacy is a personal judgment of how well or poorly a person can cope with a given situation based on the skills they have and the circumstances they face. Self-efficacy affects every area of human endeavor. By determining the beliefs a person holds regarding their power to affect situations, self-efficacy strongly influences both the power a person actually has to face challenges competently and the choices a person is most likely to make [1]. Also has been recognized that self-efficacy is a key trait of self-regulated learners [9]. Self-regulated learning refers to one's ability to understand and control one's learning environment. Self-regulation abilities include goal setting, self-monitoring, self-instruction, and self-reinforcement [9].

5 Conclusions and Future Work

The ubiquity of computers and mobile devices in classrooms, the distance education and the learning everywhere produce data every second. Educational data is rich in insights about the different aspects of students, teachers, learning processes and learning management. Therefore, machine learning techniques can be used to understand those aspects, and in turn to design and improve intelligent learning environments. A goal of educational data mining is having better educational technologies. This objective requires further advances in artificial intelligence and in human learning theories. Educational data mining is an emerging discipline that can be useful towards these aims due to its potential to support the development of fields related to education.

In this paper, we present a brief review of relevant research in educational data mining, an also we propose the modeling and designing of intelligent tutoring systems based on a data-driven approach. The ITS are being built considering the knowledge produced by the educational data mining techniques. We are analyzing public educational datasets, but also, we are gathering data by means of a controlled experiments with an online course and under graduated and graduated students participating. Despite, this research is in an initial stage, we visualize potential in the educational data mining techniques based on our previous work in intelligent learning environments.

References

1. Bandura, A.: Guide for constructing self-efficacy scales. Self-efficacy beliefs Adolesc. pp. 307–337 (2006). https://doi.org/10.1017/CBO9781107415324.004

2. Cerezo, R. et al.: Process mining for self-regulated learning assessment in e-learning. J. Comput. High. Educ. 32, 1, pp. 74–88 (2020). https://doi.org/10.1007/s12528-019-09225-y
3. Ferreira-Mello, R. et al.: Text mining in education. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9, 6, (2019). https://doi.org/10.1002/widm.1332
4. Fischer, C. et al.: Mining Big Data in Education: Affordances and Challenges. Rev. Res. Educ. 44, 1, pp. 130–160 (2020). https://doi.org/10.3102/0091732X20903304
5. Hernández, Y. et al.: Data-driven construction of a student model using bayesian networks in an electrical domain. In: 16th Mexican International Conference on Artificial Intelligence, MICAI 2017. pp. 481–490 Springer Verlag (2017). https://doi.org/10.1007/978-3-319-62428-0_39
6. Hung, J.L. et al.: Improving Predictive Modeling for At-Risk Student Identification: A Multistage Approach. IEEE Trans. Learn. Technol. 12, 2, pp. 148–157 (2019). https://doi.org/10.1109/TLT.2019.2911072
7. Koedinger, K.R. et al.: New Potentials for Data-Driven Intelligent Tutoring System Development and Optimization. AI Mag. 3, pp. 27–41 (2013)
8. Merceron, A. et al.: Learning Analytics: From Big Data to Meaningful Data. J. Learn. Anal. 2, 3, pp. 4–8 (2016). https://doi.org/10.18608/jla.2015.23.2
9. Panadero, E.: A review of self-regulated learning: Six models and four directions for research. Front. Psychol. 8, APR, 1–28 (2017). https://doi.org/ 10.3389/fpsyg.2017.00422
10. Romero, C., Ventura, S.: Educational data mining and learning analytics: An updated survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 10, 3, pp. 1–21 (2020). https://doi.org/10.1002/widm.1355
11. Witten, I.H. et al.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Cambridge (2017)
12. Woolf, B.P.: Student modeling. In: Nkambou, R. et al. (eds.) Studies in Computational Intelligence. pp. 267–279 Springer (2010). https://doi.org/10.1007/978-3-642-14363-2_13