ResearchLEAP

Journal of International Business Research and Marketing
Volume 6, Issue 1, 2020
journal homepage: www.researchleap.com

Benefits of Educational Data Mining

Alisa Bilal Zorić
Polytechnic Baltazar Zaprešić, Zaprešić, Croatia

DOI: *10.18775/jibrm.1849-8558.2015.61.3002*
URL: *http://dx.doi.org/10.18775/jibrm.1849-8558.2015.61.3002*

Keywords:
Data, Education, Educational data mining, Teaching, Students

ABSTRACT

We live in a world where we collect huge amounts of data, but if this data is not further analyzed, it remains only huge amounts of data. With new methods and techniques, we can use this data, analyze it and get a great advantage. The perfect method for this is data mining. Data mining is the process of extracting hidden and useful information and patterns from large data sets. Its application in various areas such as finance, telecommunications, healthcare, sales marketing, banking, etc. is already well known. In this paper, we want to introduce special use of data mining in education, called educational data mining. Educational Data Mining (EDM) is an interdisciplinary research area created as the application of data mining in the educational field. It uses different methods and techniques from machine learning, statistics, data mining and data analysis, to analyze data collected during teaching and learning. Educational Data Mining is the process of raw data transformation from large educational databases to useful and meaningful information which can be used for a better understanding of students and their learning conditions, improving teaching support as well as for decision making in educational systems.The goal of this paper is to introduce educational data mining and to present its application and benefits.

1. Introduction

Today, universities operate in a highly competitive and complex environment. With rapid technology development and cheaper IT equipment, the amount of data stored in educational databases increases rapidly, but if this data is not further analyzed, it remains only huge amounts of data. Data mining tools, methods and techniques, allow us to analyze this data and find hidden patterns and information. Data mining is used to detect patterns and relationships in data to improve decision-making processes. It is an interdisciplinary area that brings together techniques from statistics, artificial intelligence, neural networks, database systems, machine learning, pattern recognition, data visualization, knowledge acquisition and information theory (Sumathi and Sivanandam, 2013).

The application of data mining is wide and various. It is used in finance for analyzing customer behavior data to increase customer loyalty, it also helps in finding hidden correlations between various financial indicators to detect suspicious activities. By collecting historical data and turning them into useful and valid information it can detect fraudulent and non-fraudulent actions. Using data mining in healthcare can help in discovering the relationships between diseases and the effectiveness of treatments. It also supports healthcare insurers in detecting fraud. It is used in crime agencies for finding patterns related to money laundering, narcotics trafficking, etc.

A common use of data mining in telecommunication is in analyzing customer data to improve profitability by providing customized services and also to reduce customer churn by understanding demographic characteristics and predicting customer behavior. The results of the data mining process can be used to develop appropriate marketing campaigns and pricing strategies. In marketing and sales, data mining techniques are used to find the hidden patterns from historical purchasing data. Results of data mining provide information on combinations of products purchased together in market basket analysis and are used to identify customer's behavior buying patterns. It is also used for the prediction of future trends and customer purchase habits. The banking industry usually uses data mining methods to predict customer churn, as well as in fraud and bankruptcy detection.

There are also disadvantages of data mining, namely in user privacy and security. It has to be clear how and with whom the information will be used and shared. Data mining tools and techniques work with very big amounts of data, so there is great cost at the implementation stage. It requires great IT experts for preprocessing data and finding the best model and technique for analysis. The techniques of data mining are not 100% accurate, so it may cause serious consequences and expenses.

This work is based on special use of data mining algorithms, techniques and concepts in the educational environment, called educational data mining (EDM). The remaining of the paper is organized as follows. After a summary of the history and definition of educational data mining, the process is presented in "Educational data mining process" section, by detailing the data pre-processing and the knowledge extraction phase, and by describing all phases. After that, selected methods and techniques, as well as its use in the educational sector are described in "Methods and techniques" section. In the next section, related work is covered and in the most relevant section, benefits and applications of educational data mining are presented and discussed, along with relevant research in the application of educational data mining. Final remarks conclude into "Conclusion" section.

2. Educational Data Mining (EDM)

The amount of data collected and stored in many educational institutions grew too big and educational data analysis could not be performed manually anymore. EDM is a relatively new discipline that emerged from the application of data mining techniques on educational data. The first international research conference on EDM was in Montreal, Canada in 2008. Journal of Educational Data Mining started publishing in 2009 and the International Educational Data Mining Society was founded in 2011 (Romero and Ventura, 2010, pp. 601-618). From that point, EDM continues to grow from different research areas such as data mining and machine learning, pattern recognition, psychometrics and other areas of statistics, artificial intelligence, information visualization and computational modeling.

The final goal of EDM is to improve the educational process and to explain educational strategies for better decision making (Silva, Fonseca, 2017, p. 87). There are different definitions of EDM, but they all have common that it is an interdisciplinary research area which uses different methods and techniques from machine learning, statistics, data mining and data analysis, to analyze data collected during teaching and learning in order to discover previously unknown information, relationships and patterns in large data repositories.

2.1 Educational Data Mining Process

The EDM process has four main phases. Problem definition is the first phase in which a specific problem is translated into a data mining problem. In this phase, the project goal and objectives are formulated, as well as the main research questions. The most time-consuming phase is the second phase, Data preparation and gathering phase. It can take up to 80 % of all analysis time. Data quality is a major challenge in data mining (Blake and Mangiameli, 2011). In this phase, source data must be identified, cleaned and formated in prespecified format. After that, there is a Modeling and Evaluating phase in which the parameters are set to optimal values and different modeling techniques are selected and applied. The deployment phase is the last phase in which the results of data mining are organized and presented through graphs and reports. It is important to point out that the data mining process is an iterative process which means that the process does not stop when a particular solution is deployed. It can be just a new input for a new data mining process.

2.2 Methods and Techniques

Various methods, algorithms and techniques are used for educational data mining. The most often use is for classification, clustering, prediction and association. The most frequently used data mining techniques are neural networks, decision trees, regression analysis and cluster analysis.

Classification is a data mining technique that segments data in a collection to target categories or classes. It helps in analyzing data and predicting outcomes. The goal of classification is to accurately predict the target class for each case in the data. The classifier training algorithm uses pre-classified examples for determining the set of parameters required for classification (Oracle, 2019). In the educational sector, this technique is often used for classifying students based on some characteristics such as age, gender, grades, knowledge, academic achievements, motivation, behavior, demographic or geographic characteristics, etc.

Clustering analysis is used to segment similar data into clusters that were not previously defined. It is useful in the data-preprocessing phase to identify homogeneous groups used as input for other models. Similar to classification, cluster analysis can be used to investigate similarities and differences between students, courses, teachers, etc.

Prediction refers to calculated assumptions for certain events made based on available processed data. Regression technique can be used for prediction, to model the relationship between one or more independent variables and dependent variables. Independent variables are attributes already known and response variables are what we want to predict (Bhatnagar, 2013). It has many applications in business planning, trend analysis, financial forecasting, time series prediction, trend analysis, etc. In the educational sector, it is used for the prediction of students' academic performance, prediction of enrolled students, prediction of the final grade, prediction of drop-outs, etc.

Association is a data mining technique used to discover the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are called association rules. Association rules are frequently used to analyze sales transactions. This type of finding is valuable for sales promotions, direct marketing, catalog design, cross-sell marketing and for discovering business trends. Based on certain rules, this technique can be used for the introduction of new courses or to open new colleges.

Neural networks are a set of computational algorithms, inspired by the human central nervous system, that is designed to recognize complex patterns and prediction problems by considering learning material, without being programmed (Graupe, 2013). They automatically identify special characteristics from the examples that they process. Neural networks consist of connected nodes called artificial neurons. Every connection can transmit a signal from one to another artificial neuron. A signal is a real number. Artificial neurons and connections have weights that adjust during the learning process. Neurons are divided into three layers, input, output and hidden layer. Signals travel from the input layer through hidden layers to the output layer, performing different kinds of transformations on their inputs. Its most important ability is to learn and model non-linear and complex relationships. The most common use of artificial neural networks is for speech and image recognition, computer vision, machine translation, for playing video games, financial analysis, social network filtering, control, optimization and medical diagnosis.

A decision tree is a decision support tool that uses a tree-shaped graph or model for classification. It is a supervised learning method. Each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class which is a decision after computing all attributes. The paths from the root to leaf are classification rules (Vidal et al, 2014). Their greatest advantage is stability and easy interpretation. Because of their simplicity, they are suitable for solving a different kind of problems in a broad range of industries such as financial, business, healthcare, education, energy, engineering, pharmaceutical, law, etc.

3. Related Work

Educational data mining is a young research area which is becoming increasingly popular due to its potential. Educational data can be used to assist instructors, to improve curriculums, to understand students' behavior, to improve teaching process, to improve e-learning systems, to identify reasons for dropping out, to support decision making, etc. (Romero and Ventura, 2010). Educational data mining research can be divided into two main categories, one regarding the analysis of learning behavior and attributes that affect a successful study, and the main goal of other research is to find a predictive model for student's performance.

A review of research literature on EDM between 1995 and 2005 is covered in Educational Data Mining: A Survey from 1995 to 2005. wrote by Romero and Ventura (2007). They discussed the use of web mining techniques in education systems and compared the traditional classroom teaching and web-based education.

A systematic review of the published EDM literature between 2006 and 2013, based on the highly cited paper collected through Google scholar index, is provided by Al-Razgan and Al-Khalifa (2014).

In Educational Data Mining Applications and Trends, author gave a bibliographic review of the various educational data mining studies, used techniques and contribution to their application (Peria-Ayala, 2013).

Kumar and Chadha (2011) presented an empirical study of the applications of data mining techniques in higher education in which they tried to identify the potential areas in which data mining techniques could be applied. They concluded that potential applications are: organization of syllabus, predicting the registration of students in an educational program, predicting student performance, detecting cheating in the online examination and identifying abnormal or erroneous values and used data mining techniques are: association analysis, classification, prediction, clustering and outlier analysis.

Ali (2013) emphasized following benefits of educational data mining: identifying students' pattern trends, preferences and course needs, selection of specialization, predicting students' final results, automatic exploration of data and profiling students.

4. Benefits and Applications of Educational Data Mining

Benefits and applications of educational data mining are numerous. There are many research papers and studies regarding the use and applications of data mining techniques in education. Some of them will be described later. The most common use of educational data mining is: improving the process of studying, improving course completion, supporting students in course selection, students' profiling, finding problems leading to dropping out, students' targeting, curriculum development, predicting student's performance and as a support for decision-making at student enrolment.

Romero and Ventura (2010) pointed out these areas of application of EDM: student's modeling, predicting student's performance, data visualization, social network analysis, feedback for support management, planning and scheduling, grouping students, detection of undesirable behaviors.

In Mining educational data to analyze students' performance, Baradway and Pal (2012) pointed out the capabilities of data mining techniques in the context of higher education by offering a data mining model for the higher education system. The decision tree method was used to evaluate a student's performance. This research could help educators in the early detection of dropouts and students who need special attention to provide appropriate advising or counseling.

In second research, Data Mining: A prediction for performance improvement using classification Bhardwaj and Pal (2012) used Bayes classification for the construction of a prediction model

to identify the difference between high learners and slow learners.

Pal (2012) used the classification task to evaluate previous years of student dropout data to find students who are likely to drop out of their first year of engineering. Decision tree methods (ID3, C4.5, CART and ADT) were applied and information about previous education, student's family income, parents' education, etc. were used to predict the list of students who need special attention to reduce the drop-out rate. Results from this research show that the machine learning algorithm can establish an effective predictive model from the existing student dropout data.

Luan (2002) discussed the potential applications of data mining and explained how data mining techniques can save resources and maximize efficiency in higher education.

Kovačić (2012) examined the socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course program and course block) that may help in identifying successful and unsuccessful students. This research concluded that classifying students based on pre-enrolment information can help to identify students at-risk of dropping the course and suggest advising and mentoring programs to make them successful.

Kabackijeva (2012) applied different data mining methods (rule learner, a decision tree classifier, a neural network and the nearest neighbor classifier) to develop an enrollment prediction models based on student's personal, pre-university and university characteristics.

Maqsood (2013) stated that data mining can be used to report and analyze the data that can help in preparing marketing strategies for targeted students.

Kardan et al. (2013) focused on identifying various factors influencing student's online course selection using neural networks and applying these factors to predict the final number of students in every course.

Guo (2010) used neural networks to analyze and predict students' course satisfaction. This study showed that the most influential factors to student course satisfaction are the number of enrolled students in a course and the high distinction rate in the final grading.

Mardikyan and Badur (2012) used two different data mining techniques: stepwise regression and decision trees to identify varied factors that affect instructors' teaching performance in university.

Tsai et al. (2011) applied three different cluster techniques (k-means, self-organizing maps, and two-step clustering) to cluster students into different groups based on their computer literacy. After clustering students into different groups, the decision tree algorithm was used to extract useful rules from each group. This research concludes that data mining techniques can help universities to identify several groups who need additional training to pass a computer proficiency test.

Hsia et al. (2008) used different data mining algorithms: decision tree, link analysis and decision forest to study course preferences, completion rates and profession of enrollees. The results of this study found the correlation between course category and enrolee professions and pointed out the importance of data mining in building curriculum and marketing in the field of higher education. These results may be used as a reference for marketing and curriculum development.

Understanding students' behavior and how they learn can help educational management to improve current study programs and educational practice in general. By analyzing the educational data, as well as analyzing the importance of the influence of individual variables, various data mining models could be used as support for decision-making in education, thus contributing to a more successful study and enhancing the quality of education. Educational data mining results can help universities to allocate resources more effectively.

5. Conclusion

Educational data mining is a young discipline with high potential for every participant in the educational process. Data mining techniques were developed to automatically discover hidden knowledge and recognize patterns from data. Educational data mining can be used for classifying and predicting students' performance, dropouts as well as teachers' performance. It can help educators to track academic progress to improve the teaching process, it can help students in course selection and educational management to be more efficient and effective.

Educational data mining can be used to attract, maintain and retain the students to achieve the profitability of University. Analyzing students' data is crucial for discovering, detecting and understanding which instructional practices are effective. In this paper, we presented the benefits and applications of data mining techniques in many educational areas.

The main goal of the paper is to reveal the high potential of educational data mining applications and to encourage others to use it.

References

- Al-Razgan, M., Al-Khalifa, A. S., Al-Khalifa, H. S. (2014). Educational data mining: A systematic review of the published literature 2006-2013. In Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013) (711-719). Singapore: Springer. Crossref

- Ali, M. M. (2013). Role of data mining in education sector. International Journal of Computer Science and Mobile Computing, 2(4), 374-383.
- Baker, R. S., Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM| Journal of Educational Data Mining, 1(1), 3-17.
- Baradwaj, B. K., Pal, S. (2012). Mining educational data to analyze students' performance. Retrived 25.5.2019. from https://arxiv.org/pdf/1201.3417.pdf.
- Bhardwaj, B. K., Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. Retrived 25.5.2019. from https://arxiv.org/pdf/1201.3418.pdf.
- Bhatnagar, V. (Ed.). (2013). Data mining in dynamic social networks and fuzzy systems. IGI global. Crossref
- Blake, R., Mangiameli, P. (2011). The effects and interactions of data quality and problem complexity on classification. Journal of Data and Information Quality (JDIQ), 2(2), 8. Crossref
- Daniel, G. (2013). Principles of artificial neural networks (Vol. 7). World Scientific. Retrived 15.07.2019. from https://books.google.hr/books?id=W6W6CgAAQBAJ&printsec=frontcover&redir_esc=y#v=onepage&q&f=false
- Guo, W. W. (2010). Incorporating statistical and neural network approaches for student course satisfaction analysis and prediction. Expert Systems with Applications, 37(4), 3358-3365. Crossref
- Han, J., Pei, J., Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
- Hsia, T. C., Shie, A. J., Chen, L. C. (2008). Course planning of extension education to meet market demand by using data mining techniques–an example of Chinkuo technology university in Taiwan. Expert Systems with Applications, 34(1), 596-602. Crossref
- Kabakchieva, D. (2012). Student performance prediction by using data mining classification algorithms. International journal of computer science and management research, 1(4), 686-690.
- Kardan, A. A., Sadeghi, H., Ghidary, S. S., Sani, M. R. F. (2013). Prediction of student course selection in online higher education institutes using neural network. Computers Education, 65, 1-11. Crossref
- Kovacic, Z. (2012). Predicting student success by mining enrolment data. Research in Higher Education, 15.
- Luan, J. (2002). Data mining and its applications in higher education. New directions for institutional research, 2002(113), 17-36. Crossref
- Maqsood, A. M. (2013). Customer Relationship Management in B-Schools: An Overview. International Journal of Computer Sciences and Management Research, 2(4), 2108-2119.
- Mardikyan, S., Badur, B. (2011). Analyzing Teaching Performance of Instructors Using Data Mining Techniques. Informatics in Education, 10(2), 245-257.
- Oracle, (2019). Retrieved 23.07.2019. from Crossref
- Peña-Ayala, A. (Ed.). (2013). Educational data mining: applications and trends (Vol. 524). Springer.
- Pena, A., Domínguez, R., Medel, J. (2009). Educational data mining: a sample of review and study case. World Journal On Educational Technology, 1(2), 118-139.
- Romero, C., Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications. Vol. 33, pp. 135-146. Crossref
- Romero, C., Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618. Crossref
- Silva, C., Fonseca, J. (2017). Educational Data Mining: a literature review. In Europe and MENA Cooperation Advances in Information and Communication Technologies (pp. 87-94). Springer, Cham. Crossref
- Sun, H. (2010). Research on student learning result system based on data mining. IJCSNS, 10(4), 203. (9)
- Tsai, C. F., Tsai, C. T., Hung, C. S., Hwang, P. S. (2011). Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. Australasian Journal of Educational Technology, 27(3). Crossref
- Vidal, J. C., Lama, M., Vázquez, B., Mucientes, M. (2014, September). Reconstructing IMS LD Units of Learning from Event Logs. In European Conference on Technology Enhanced Learning (pp. 345-358). Springer, Cham. Crossref