EasyChair Preprint

№ 9763

Tools and Methods of Educational Data Mining: a Review

Pushpita Chakrabarty, Koushik Halder and Pravakar Rao

February 22, 2023

Tools and Methods of Educational Data Mining: A Review

Pushpita Chakrabarty[1], Koushik Halder[2], Pravakar Rao[3]
Indira Gandhi Open University, India[1,2]
Jeypore Vikram Deo College of Science and Technology, India[3]

Abstract

For the goal of performing educational data mining (EDM) research, a wide range of tools have arisen in recent years. We intend to highlight some of the most popular, easily accessible, ad effective tools for researchers interested in performing EDM research in this article. We will highlight how useful these tools are for the standard data pretreatment and analysis processes in a research project, as well as other specific details like price and usability. We will also highlight specialized tools in the subject, such as text analysis, social network analysis, and data visualization tools. Finally, we'll talk about how critical it is to become familiar with a variety of tools.

Keywords: Educational Data mining, clustering methods, educational technology, data analysis tools

Introduction

The non-independence of educational data is a typical feature. For instance, we must take into account that comments are not statistically independent of one another because they may come from the same student or conversation when classifying whether they are on-topic or off-topic in education debates. Both the computation and validation of models could be harmed by this. The confined context of particular research projects and educational settings is often where the results of EDM research are attained. It is unclear how general these findings are, for example, whether the same student model parameters can be applied to diverse student groups or whether a prediction model is still accurate when applied in a new situation.

Educational data mining is an interdisciplinary field that involves the use of data analysis, machine learning, and statistical techniques to study and understand educational data. The goal of educational data mining is to extract meaningful information from large and complex educational datasets, and to use this information to improve education and support evidence-based decision making. In this review article, we will provide an overview of the tools and methods commonly used in educational data mining, and discuss their strengths and limitations.

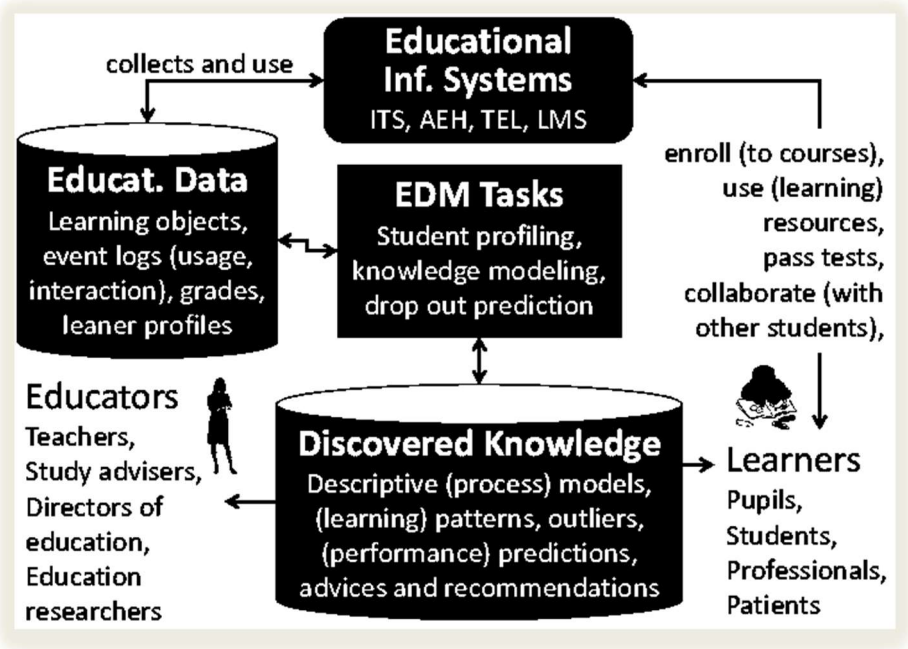

Figure 1: Framework for Educational Data Mining []

Data Preprocessing

Data preprocessing is a critical step in the educational data mining process, as it involves cleaning, transforming, and preparing the data for analysis. Common preprocessing techniques include missing value imputation, data normalization, and feature selection. There are various definitions for educational data mining. In order to be more proactive in identifying at-risk pupils and taking appropriate action, many people have turned to academic analytics, which uses statistical methods and data mining. The outcomes of data mining can be applied in this way to enhance student learning. Academic analytics is concerned with activities that take place at the departmental, organizational, or academic-university level. This form of analysis has a global viewpoint because it does not concentrate on the specifics of each particular course. A subfield of educational data mining is academic analytics.

Machine Learning Algorithms

Machine learning algorithms are a key tool in educational data mining and are used to model and predict student outcomes based on historical data. Common machine learning algorithms used in educational data mining include decision trees, random forests, support vector machines, and neural networks.

Classification and Regression

Classification and regression are two common types of machine learning algorithms used in educational data mining. Classification algorithms are used to predict categorical outcomes, such

as student success or failure, while regression algorithms are used to predict continuous outcomes, such as student grades or performance.

Clustering

Clustering is a type of unsupervised machine learning algorithm that is used to group students into homogeneous clusters based on their learning styles or behaviors. Clustering can be used to identify at-risk students, to personalize instruction, and to understand student performance.

Association Rule Mining

Association rule mining is a type of data mining technique that is used to discover relationships and patterns in educational data. Association rule mining can be used to identify factors that influence student performance, to support instructional design, and to understand student behavior. Researchers have used advanced algorithms like ant colony optimization and frequent pattern graph for the analysis of association rule [3][7].

Online Analytical Processing

Some study proposes [5][9] a design for an automated scaffolding system that generates real-time data regarding a learner's experience in a course using Online Analytical Processing (OLAP). During synchronous tutorial sessions, the scaffolding gives the instructor knowledge about the learner's behavior and self-regulated learning. The suggested scaffolding system's overall architecture creates a temporary platform for live sessions between the teacher and student (s). The scaffolding's design is more abstract, and the implementation's technical specifics will depend on the system that will be used to build it. The purpose of the study is to offer a framework for developing a more flexible and customized learning environment.

Data Visualization

Data visualization is an important tool in educational data mining, as it can be used to visually represent the results of data analysis and to communicate findings to stakeholders. Common data visualization techniques used in educational data mining include bar charts, scatter plots, and heat maps.

EDM Tools

RapidMiner: It is a toolkit for building models and carrying out data mining analysis. It offers limited capabilities for selecting features, creating multiplicative interactions, and building new features out of existing features[11].

WEKA: It is a collection of free and open-source software that includes numerous data mining and model building methods. Although it supports automatic feature selection, it does not support the introduction of new features[12].

SPSS : SPSS is well-known and not limited to the data science community, much like Excel. The main function of SPSS is statistics, and it provides a variety of statistical tests, regression models, correlations, and factor analyses. IBM SPSS Modeler Premium, a more recent analytics and data mining tool that merges earlier analytics and text mining tools, is a supplement to SPSS[13].

KNIME: A data cleaning and analysis tool formerly known as Hades, it is largely comparable to RapidMiner and WEKA. It incorporates all ML algorithms, just like RapidMiner, and provides many of the same features as other products. Additionally, it provides a variety of advanced algorithms in disciplines like sentiment analysis[14].

Tableau: A collection of interactive data analysis and visualization products are available from Tableau. The Tableau toolset's primary goal is to support business intelligence, but it has also frequently been used in educational settings to analyze student data, offer insights that can be put into practice, improve teaching methods, and streamline educational reporting[15].

Conclusion

Educational data mining is a rapidly growing field that has the potential to revolutionize education by providing data-driven insights into student learning and behavior. The tools and methods used in educational data mining, such as machine learning algorithms, clustering, and data visualization, are critical for extracting meaningful insights from educational data. However, it is important to use these tools and methods carefully, as they can also introduce biases and limitations into the analysis. Further research is needed to develop and refine the tools and methods of educational data mining, and to ensure that they are used in an ethical and effective manner.

References

[1] Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. Journal of Educational and Behavioral Statistics, 42(1), 85-106.

[2] Prabha, S. L., & Shanavas, A. M. (2014). Educational data mining applications. Operations Research and Applications: An International Journal (ORAJ), 1(1), 23-29.

[3] Sengupta, S., & Dasgupta, R. (2010). A Data Mining Approach to Determine an Efficient Learning Path. In EEE 2010: proceedings of the 2010 international conference on e-learning, e-business, enterprise information systems, & e-government (Las Vegas NV, July 12-15, 2010) (pp. 59-62).

[4] Huebner, R. A. (2013). A Survey of Educational Data-Mining Research. Research in higher education journal, 19.

[5] Sengupta, S., Mukherjee, B., Bhattacharya, S., & Dasgupta, R. (2012). OLAP based Scaffolding to support Personalized Synchronous e-Learning. International Journal of Managing Information Technology, 4(3), 71

[6] Jindal, R., & Borah, M. D. (2013). A survey on educational data mining and research trends. International Journal of Database Management Systems, 5(3), 53.

[7] Sengupta, S., Sahu, S., & Dasgupta, R. (2012). Construction of learning path using ant colony optimization from a frequent pattern graph. arXiv preprint arXiv:1201.3976.09
[8] Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. *Acm Sigkdd Explorations Newsletter*, *13*(2), 3-6.
[9] Sengupta, S., Mukherjee, B., & Bhattacharya, S. (2012). Designing a scaffolding for supporting personalized synchronous e-learning. Department of Computer Science & Information Technology, Bengal Institute of Technology, Kolkata-150, India.
[10] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.
[11] http://rapid-i.com/content/view/181/190/
[12] http://www.cs.waikato.ac.nz/ml/weka/
[13] http://www.ibm.com/analytics/us/en/technology/spss/
[14] http://www.knime.org
[15] http://www.tableau.com