

# A Review on Educational Data Mining

Ritu Gautam<sup>1</sup>, Deepika Pahuja<sup>2</sup>

<sup>1,2</sup>Lecturar, DAV Institute of Management, Faridabad, India

**Abstract:** Growing interest in data and analytics in education, teaching, and learning raises the priority for increased, high-quality research. Data Mining is a technique used to find out possibly new information from huge amount of data. Educational data mining is an emerging trend, concerned with developing methods for exploring the huge data that come from the educational system. The objective of this research is to introduce Educational Data mining, by describing a step-by-step process using a variety of techniques. In this paper a review is conducted on step by step processes and application areas.

**Keywords:** EDM, LMS, educational data mining

## 1. Introduction

Data mining is used to find out new and useful information from large amount of data. Techniques of data mining are useful in various application areas like fraud detection, businesses, banking and telecommunications. The major application of Data Mining Technique is educational data mining in order to extract useful information from educational data. In Education Quality Assurance has compelled academics for constantly exploring different ways for making improvements in educational processes. It has led to increasing interest in educational data mining. Now –a-days educational data mining tends to focus on developing new tools to discover patterns in data. The patterns are generally used to find the micro concepts involved in learning: one-digit multiplication, subtraction with carries, and so on.

Educational data mining is generally used to do research and help in building models in several areas influencing online learning systems.

Data Mining Research Framework

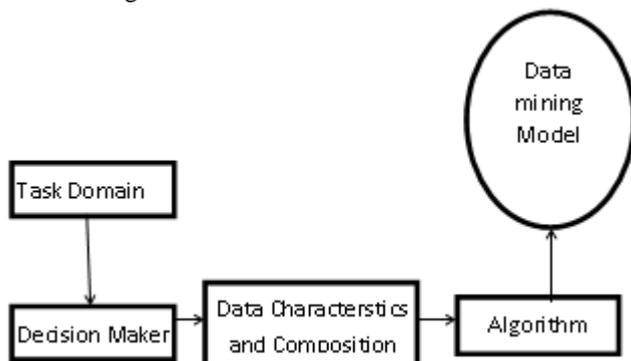


Figure 1: Data Mining Model

Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people's learning activities in educational organizations.

## 2. Illustrations

a) Learning management systems (LMSs): It is used to track information like whenever each student access a learning object, it displays no. of times they accessed it,

and for how much time the learning object on the user's computer screen.

b) Intelligent tutoring systems :It is used to record data each time a user submits a solution for a problem; it also collects the submission time, whether or not the solution matches the expected solution, the amount of time that has passed since the last submission and the order in which solution components were entered.

An important and unique feature of educational data is that they are hierarchical. Data at the keystroke level, the answer level, the session level, the student level, the classroom level, the teacher level, and the school level are nested inside one another (Baker 2011; Romero and Ventura 2010).Some Other important features are time, sequence, and context. Time is crucial important to capture data, such as length of practice sessions or time to learn. Sequence represents how concepts build on one another and how practice and tutoring should be ordered. Context is important for explaining results and knowing where a model may or may not work.

Some of the important developments in mining educational data are hierarchical data mining and longitudinal data modeling are introduced.

## 3. Phases in Educational Data Mining

Educational data mining is the emerging area in the field of research and continuously growing, a v of variety of mining techniques has been accomplished in a variety of educational contexts. The major objective is to translate raw data into meaningful information about the learning process to make better decisions about the design and trajectory of a learning environment. EDM generally consists of four phases:

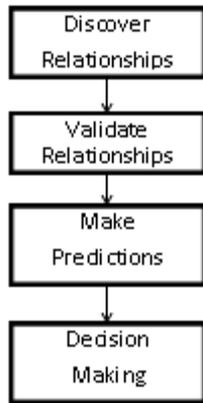


Figure 2: Phases in Educational Data Mining

#### 4. Discover Relationships

The first phase in the EDM process (not including pre-processing) is to discover relationships between data. It is accomplished by searching in a repository of data in an educational environment with the objective of finding consistent relationships between data variables.

Various algorithms are used for finding such relationships some of them are classification, regression, clustering, factor analysis, social network analysis, association rule mining, and sequential pattern mining.

##### 1) Validating Relationships

Discovered relationships must be validated in order to avoid over fitting.

##### 2) Making Predictions

Relationships which are valid are used to make predictions about further events in the learning environment.

##### 3) Decision Making

Predictions made in previous phase are used to support decision-making processes and in making policy decisions. During phases 3 and 4, data is often visualized or in some other way to make distillation of human judgment.[2] Visualising Data, is one of the best practices in research.

##### 4) Goals

Baker and Yacfi identified the following four goals of EDM:

- a) **Predicting students' future learning behavior**– With the help of student modeling, the goal can be achieved by creating student models which incorporates the learner's characteristics that includes the detailed information such as their knowledge, behaviours and motivation to learn. The experience of the user and learner with their overall satisfaction while measuring learning.
- b) **Discovering or improving domain models** – with the help of methods and applications of EDM, introduction of new model and improvements in existing models. Learners are engaged in educational content in determining optimal instructional sequences in order to support the learning style of students.
- c) **Studying the effects of educational support**- It can be achieved with the help of learning systems.

d) **Advancing scientific knowledge about learning and learners**- It is done by building and incorporating student models, technology and software are also used in the field of EDM research.

e) **Pedagogy**- It helps in Studying the effects of different kinds of pedagogical support that can be provided by learning software and Advancing scientific knowledge about learning and learners through building computational models which incorporates models of the student, the domain, and the software's pedagogy[1].

#### 5) Technical Methods

To accomplish these goals, educational data mining researchers use the five categories of technical methods (Baker 2011) described below.

- a) **Prediction** – It entails in developing a model that can infer a single aspect of the data (predicted variable) from the combination of other aspects of the data (predictor variables). Examples of using prediction include detecting such student behaviors as when they are gaming the system, engaging in off-task behavior, or failing to answer a question correctly despite having a skill. Predictive models have been used for understanding what behaviors in an online learning environment—participation in discussion forums, taking practice tests and the like—will predict which students might fail a class. Prediction shows promise in developing domain models, such as connecting procedures or facts with the specific sequence and amount of practice items that best teach them, and forecasting and understanding student educational outcomes, such as success on posttests after tutoring (Baker, Gowda, and Corbett 2011).
- b) **Clustering** refers to finding data points that naturally group together and can be used to split a full dataset into categories. Examples of clustering applications are grouping students based on their learning difficulties and interaction patterns, such as how and how much they use tools in a learning management system (Amershi and Conati 2009), and grouping users for purposes of recommending actions and resources to similar users. Data as varied as online learning resources, student cognitive interviews, and postings in discussion forums can be analyzed using techniques for working with unstructured data to extract characteristics of the data and then clustering the results. Clustering can be used in any domain that involves classifying, even to determine how much collaboration users exhibit based on postings in discussion forums (Anaya and Boticario 2009).
- c) **Relationship mining** involves discovering relationships between variables in a dataset and encoding them as rules for later use. For example, relationship mining can identify the relationships among products purchased in online shopping (Romero and Ventura 2010).
  - *Association rule mining* can be used for finding student mistakes that co-occur, associating content with user types to build recommendations for content that is likely to be interesting, or for making changes to teaching approaches (e.g., Merceron and Yacfi 2010). These techniques can

be used to associate student activity, in a learning management system or discussion forums, with student grades or to investigate such questions as why students' use of practice tests decreases over a semester of study.

- *Sequential pattern mining* builds rules that capture the connections between occurrences of sequential events, for example, finding temporal sequences, such as student mistakes followed by help seeking. This could be used to detect events, such as students regressing to making errors in mechanics when they are writing with more complex and critical thinking techniques, and to analyze interactions in online discussion forums.

Key educational applications of relationship mining include discovery of associations between student performance and course sequences and discovering which pedagogical strategies lead to more effective or robust learning. This latter area—called teaching analytics—is of growing importance and is intended to help researchers build automated systems that model how effective teachers operate by mining their use of educational systems.

- 6) Distillation for human judgment is a technique that involves depicting data in a way that enables a human to quickly identify or classify features of the data. This area of educational data mining improves machine-learning models because humans can identify patterns in, or features of, student learning actions, student behaviors, or data involving collaboration among students. This approach overlaps with visual data analytics (described in the third part of this section).
- 7) Discovery with models is a technique that involves using a validated model of a phenomenon (developed through prediction, clustering, or manual knowledge engineering) as a component in further analysis. For example, Jeong and Biswas (2008) built models that categorized student activity from basic behavior data: students' interactions with a game-like learning environment that uses learning by teaching. A sample student activity discerned from the data was "map probing." A model of map probing then was used within a second model of learning strategies and helped researchers study how the strategy varied across different experimental states. Discovery with models supports discovery of relationships between student behaviors and student characteristics or contextual variables, analysis of research questions across a wide variety of contexts, and integration of psychometric modeling frameworks into machine-learned models.

### 5. Applications of EDM Methods

There have been a wide number of applications of educational data mining. The four areas of applications that have received the particular attentions, including individual learning from educational software, collaborative learning, computer-adaptive testing and analyzing students' performance. One key area of applications is in improving student models. The model provides detailed information about a student's characteristics or states, such as

knowledge motivation and attitude. The educational data mining methods have enabled researchers to make higher-level inferences about students' behavior, such as when a student is gaming the system [5], and when a student is engaging in self-explanation [40]. These students' models have increased the instructors' ability to predict the students' knowledge and future performance have enabled researchers to study what factors lead students to make specific choices in a learning setting. A second key area of application is in discovering or improving models of the knowledge structure of the domain. In educational data mining, methods have been created for discovering accurate domain models directly from data. The combinations of psychometric modeling frameworks with advanced space-searching algorithms are posed as a prediction problem for the model discovery. Barnes [7] has developed algorithms which can automatically discover a Q-matrix from data. Pavlik et al. [31] has proposed algorithms for finding partial order knowledge structure models by looking at the covariance of individual item sets.

A third key area of application of EDM methods has been in studying the pedagogical support provided by learning software. Educational data miners are interested in discovering most effective pedagogical support. Learning decomposition fits exponential learning curves to performance data, relating students' success to the amount of each type of pedagogical support a student has received [8]. A fourth key area of application of educational data mining is for specific discovery about learning and learners. Discovery with models is a key method for scientific discovery via Educational data mining. Perera et al. [32] used the big five theory for teamwork as a driving theory to search for successful patterns of interaction within student teams.

**Table 1**

<i>Application Area</i>	<i>Questions</i>	<i>Type of Data needed</i>
User Knowledge Modelling	About Specific Skills and procedural concepts.	Students Response ,Time And Changes while investigation period
User Experience Modelling	About user Satisfaction	Response to survey, choices and questionnaires etc.
User Profiling	What group of users make clusters	Students response, time, hints, errors repetition.
Domain Modelling	About level correctness, module division and sequencing of modules	Performance of modules ,module taxonomy and associations
Learning component analysis and instructional principle analysis	Which are effective components, learning principle and effective curriculum	Modules performance at different levels compared to external measures.

### 6. Limitation of Data Mining

Data mining is still an art which requires skillful analysis and methodology's choice. It requires expertise in subject area, experience in handling large data bases and data mining algorithms. The techniques being used with the help of existing classical statistics or artificial intelligence techniques. Problems discovered while data includes over-

fitting existing data, missing and noisy data, and dealing with very large databases and very high dimensionality need to be resolved.

Furthermore, for large-scale, real-world tasks, high performing algorithms such as neural networks and genetic algorithms must cope with long computation times and difficulties in making interpretations. Techniques associated with probabilistic learning requires to be improved.

Data mining and analytics in classrooms, schools, districts, and other institutions to be successful. This will enable them to pose questions that matter to teachers and other users and to frame findings in a thoughtful, informative way that highlights and recommends clear actions. In reports about the newest technologies for adaptation, personalization, and recommendation, the role of human judgment is sometimes underemphasized (with the exception of visual data analytics). All the experts consulted for this issue brief emphasized the key role that people play in many steps of the data mining and analytics process. Smart data consumers can help determine what questions to address, what data to collect, and how to make reports meaningful and actionable. They can also help interpret data, discern and label patterns, and guide model building. Data mining and analytics technology play a supporting role in the essentially human and social effort of making meaning out of experience. One expert interviewed stressed that data mining and analytics do not give answers when just unleashed on a big data warehouse. Instead, the recommendation was to approach the problem in an informed way, considering what can be acted on, what evidence can come from data analysis, and what early pilots of the data mining and analytics applications reveal. Smart data consumers must learn to keep an open mind to what the data say. Data mining and analytics techniques can confirm or disconfirm teachers' and students' beliefs about student knowledge, abilities, and effort. Sometimes, these beliefs are not consistent with the data: Teachers may believe particular students are more or less capable than they are, and students may report spending more time and effort on learning than they actually do. For example, one company found in an A/B study it conducted on the use of visualizations that students were more engaged when complex visualizations were included in the software. Students identified complexity as a source of their engagement, but teachers thought the visualizations were too complex, underestimating what the students were capable of understanding.

## 7. Conclusion

This paper concludes that educational data mining is used in various application areas with the help of different methods and proposing the method that can check the correctness of forms filled by students for university using domain clustering.

## References

- [1] EdTechReview™ (ETR).htm
- [2] International Journal of Computer Science, Engineering and Information Technology (IJCSIEIT), Vol.2, No.3, June 2012
- [3] Baker, R.S.J.d., Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1 (1), 3-17.
- [4] International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014. www.ijiset.com ISSN 2348 – 7968
- [5] IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012 ISSN (Online): 1694-0814 www.IJCSI.org
- [6] International Conference & Workshop on Recent Trends in Technology, (TCET) 2012 Proceedings published in International Journal of Computer
- [7] Baker, R. S. J. d. 2011. "Data Mining for Education." In International Encyclopedia of Education, 3<sup>rd</sup> ed., edited by B. McGaw, P. Peterson, and E. Baker. Oxford, UK: Elsevier.
- [8] U.S. Department of Education. 2010a. National Education Technology Plan. <http://www.ed.gov/technology/netp-2010>.
- [9] Asia pacific journal of research October 2013, Volume: I, Issue: X
- [10] MCSER-Mediterranean Center of Social and Educational Research Rome, Italy, 2011 [www.mcser.org](http://www.mcser.org)
- [11] Dunham, M. (2003). Data Mining: Introductory and Advanced Topics Pearson Education.