Educational Data Mining: A Review

P. Bachhal¹, S. Ahuja¹ and S. Gargrish¹
¹Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India.

Email: prabhnoor.bachhal@chitkara.edu.in

Abstract: Data mining in the field of education plays a vital role. The main objective of this analysis is to understand how researchers have done data mining in the past and the current data mining developments in educational research. It describes how academic data and learning analytics applied to academic data. EDM uses computational methods to evaluate data concerning education to research questions related to education. This paper discusses the most important studies conducted to date in this area. EDM is implemented and the various user groups, various educational environments/data are defined. The problem solved by data mining techniques in different areas to improve the success of students is addressed.

Keywords: Educational Data Mining, Educational Settings, Knowledge discovery, Data mining.

1. Introduction to Educational Data Mining

Data mining is the extraction of important (non-trivial, implicit, previously unknown and potentially useful) information or trends from large amounts of data [1]. To gather large amount of information from data is called data mining. Earlier, this work was supposed to be done by data analyst but now computers have changed this and it is more efficient than statistical method. In broad data sets, data mining is the method of finding useful patterns and trends. There are various tasks in data mining like description, prediction, estimation, classification, clustering and association. Data mining is a method that businesses use to convert the raw data into valuable information. Data mining depends on effective data gathering, storage and computer processing [3].

Data mining applications plays an important role in the education system. However, data mining techniques helps in improving performance and efficiency. Computer education system face many problems like while admission process, it is hard to predict the performance of each student that if he/she will qualify this course or not because researchers did used the conventional method for educational data base and though in today’s world, the use of data mining in education system becomes more popular which eventually provided a new name in education system called as educational data mining [6]. In educational data mining (EDM), only useful information can be extracted from large data base. It is used to predict the future performance of the students [1,2,7,8]. With EDM’s help, result of the student’s performance will be improved by performing different activities. In education, Internet plays a vital role and help students in studies known as e-learning or web learning where large amount of data is globally available and helps in improving their performance [2]. Two main applications of data mining have been used in education system i.e. Education data mining and learning analytics [1].

EDM is concerned with the development of methods for exploring unique data types to better understand students and the contexts in which they learn in educational settings [9]. On one hand, the internet creates a new history in digital learning where vast amount of knowledge about the
relationship between teaching and learning is endlessly developed and globally available [3]. On the other hand, growth of both educational software and database systems of student’s knowledge has set up massive data stores reflecting how learners think [5]. All this kind of information offers a valuable source of knowledge and skills [4].

2. Goals of Educational Data Mining

The four good EDM goals are given by Baker and Yacef [66] as shown in figure 1.

A. Predicting future learning patterns for students:
By considering the actions of the student, this objective can be accomplished by developing student models which assimilate the characteristics of the learner, including detailed information such as their awareness, attitudes, intelligence and learning motivation.

B. Invention/improvement of domain models:
By using different approaches and techniques in EDM applications, development of alternatives and enhancements to existing models can be achieved. Through integrating psychometric modelling systems with space-searching algorithms a number of researchers were able to create automatic approaches from the literature of machine learning that can find accurate models of domain structure, directly from information.

C. Advancing scientific knowledge of students and instructors:
Designing and integrating student templates, using EDM analysis and the tools and applications used through, knowledge creation can be accomplished.

D. Studying the impact of learning support:
EDM approaches have been searching for empirical data to refine and broaden educational hypotheses and well-known educational patterns to gain a deeper understanding of the key factors that influence learning, also in order to build improved learning systems.

3. Objective for using data mining in education field:

In educational data mining, it has both applied and pure research objectives i.e. to improve the learning process and guiding learning for the students whereas pure research is used to achieve a deep understanding of phenomena in education.

Data- There is several different forms of data available for mining in the educational environments. These data are unique to the field of education and therefore have inherent semantic knowledge, interactions with additional content.

Techniques- Academic data contains several unique features that involve treatment of the mining problem in a different way. While it is possible to explicitly apply most conventional DM methods, some cannot and must be tailored to the particular educational issue at hand.

Educational data mining comprises multiple user or participant classes. In accordance with their own mission, vision and priorities for the use of data mining, various groups look at educational knowledge from different angles [59]. Information revealed by EDM algorithms, for instance, will not only help teachers control their classes, recognize their classes, and the students grasping power and their own teaching approaches, but also to encourage the reflection of a learner on the situation and give learners feedback [58]. Even though, an initial concern appears to include only two main types, the students/pupils and the teachers, as can be seen in Table 1.
Figure 1. Goals of EDM.

4. Educational Settings and Types of Data/Environment:
In recent years, EDM has emerged as a research field for researchers from diverse and related research areas around the world, like physical (offline) learning, digital learning and the management system of learning and Blended Learning. There are many kinds of data accessible for extraction in the education settings as shown in fig 2. These data are special to the field of education. EDM worked with the creation of methods for assessing and using various types of data from educational settings to better understand students and their learning environments.

Table 1. EDM users.

| Students | Personalizing e-learning; proposing programmes for students and tools and learning tasks to enhance their learning skills; suggesting to the student’s enthusiastic interactions of learning; Suggested plucking and shortening or simply following links, creating adaptive tips, suggesting courses, related discussions, books, etc. |
| Instructors/Teachers | To get objective suggestions about instruction; The learning and actions of the students; Predicting academic performance. |
| Pedagogical Experts/ Educational Researchers | To review and preserve the courseware; to enhance the learning of students; Assessing the quality of the course layout and its value in the course of learning; build styles for students and tutorial models automatically; Data mining methods to suggest the most effective for each mission. Developing innovative data mining instruments for educational purposes. |
| System Administrators | Find the right way for institutional collaboration (Social and practical) resources and their academic offers; make more efficient use of available resources; Develop educational offerings and assess the efficacy of the distance learning approach. |
| Private Training Companies/ Organizations/Firms | Improving decision-making processes at higher education institutions; to meet particular goals; to recommend such courses that could be of benefit to each class students. Choosing other very talented eligible graduation applicants; to help to accept students who will be doing well in universities. |
Educational data mining encourages new knowledge based on data to be found used by the students in order to assist validation/evaluation of education programs [10]. Some similar thoughts have already been successfully incorporated in e-commerce systems, the first and most common data mining application, evaluating customer expectations with a view to rising retail purchases [11]. There are several important issues which distinguish DM application especially to education from how it is used in other fields [12].

Figure 2. Educational Data Mining.

4.1 Traditional learning / Face-to-face learning:
Offline education is attempting to communicate information and skills based upon face-to-face communication as well as studies physically and mentally about how people learn. Psychometric and analytical methods were applied to data such as student attitude/achievement, education etc. gathered in class activities. Most widely used educational settings environment is traditional classrooms. It is based on a face-to-face interaction, coordinated through instructors, between teachers and learners. In traditional classrooms, teachers try to improve instruction by observing the learning processes of students and by reviewing their performance through paper documents and assessment. They can also use attendance system information, course details, and curriculum objectives and individualised plan data. Universities also want to know which students are going to participate in a specific course, and which students are going to need guidance having to graduate.

For information, an administrator can want to find out like the criteria for entrance and predict the size of the timetable for class enlistment. Students will get to know how best to choose courses based on how well they are going to do in the chosen courses. Teachers may want to know what teaching methods apply the most to overall academic achievement, why does one class perform better the other, similar group of students. Real added advantage of offline learning is explained in table 2.

4.2 Online learning:
E-learning provides digital training, while the learning management system provides connectivity, coordination, management and monitoring instruments. Educational data mining techniques applied to data collected from students. Online education and private tutoring are methods and techniques for learners that are separated from lecturers by time and place. With access to educational programmes E-Learning programmes ignore stronger connections between students and educators (one to one).
This online based education programs will typically document accesses by the student in web logs providing a raw record of the learners’ surfing of the site. Real added advantage of online learning is explained in table 2. There are different types of logs [13]:

- **Server log file**: It comprises the data source most widely used for data mining, consisting only of mere timing, distance and input-response information.
- **Client log file**: This contains a combination of file system, one per pupil, containing details about the interaction of the user with the system.
- **Proxy log file**: It is composed of log files between user browser and server providers for caching. The data about the server log file is provided by this information.

### 4.3 Blended learning:
Blended learning is an instructional approach that incorporates digital learning resources and opportunities for online connections with conventional classroom-based strategies. It needs both teacher and student’s physical presence, including certain aspects of influence over time by students, distance, and direction/position. It is an instructional approach that incorporates digital learning resources and opportunities for online connection with conventional classroom-based strategies.

### 5. Educational tasks and Data mining techniques
There are some unique features of educational data and problems that give a particular treatment of the issue of mining. Although certain conventional DM tools should be used specifically and must be tailored to the problem’s unique learning existence. In order to solve unique educational problems, various methods of data processing can also be used.

In educational settings there are several problems or tasks which have been addressed by DM. For example, Baker [6][14] suggests 4 primary application areas for EDM: improving models for the students, to enhance Domain Models, researching digital learning educational help, learner science studies; and five methods of predicting, clustering, mining relationship, distilling data on human decision and modelling.

#### A. Data analysis and visualisation
Data analysis and visualisation are aimed at illustrating useful knowledge and helping to make decisions. For instance, it can allow teachers and course administrators to assess the learners’ course interactions and using it to obtain an awareness of the performance of the pupil in the educational system. Facts and visualisation are commonly used strategies for this interface. Statistics is a computational science of gathering, evaluating, interpreting or describing data and presenting data [15]. The obtaining of basic statistical data from statistical applications is relatively easy like SPSS. This descriptive study can be combined with educational data to include these global features of data as descriptions and reporting about the action of learners [21].

Academic data statistical analysis (log files / databases) going to tell us stuff like: the most related sites, the apps of students appear to be using, where learners join or exit using trends over time[17]; total averages of discussion site comments, number of posts vs. answers, number of interactions between learners vs. interactions between learners and teachers [16]; a student devotes a section or a portion of the course to that subject at what time[20]; visualisation of knowledge uses interactive tools to help person recognises and analyse data [18]; Online training for students, including conversations and responses to students, errors, instructor feedback, etc. [19].
Table 2. Offline vs. Online.

| Offline education/ Traditional education | Online Education |
|------------------------------------------|------------------|
| • Participants are expected to travel to the training site, usually a lecture hall, college or classroom, through offline learning. | • We are slowly and gradually moving away from the advantages that could be mined through the offline learning which is the concept of social relations which are very important aspects. Online learning is Computer or Internet based learning. Preparation can be done with online learning from virtually anywhere in the world. |
| • The role of mentor is very important. You get to learn more of values or invite the values of motivation discipline (social structure or social relations) how to maintain a social relation and students when is interacting directly knows and expects what kind of feedback, he can get and this feedback is usually a very instantaneous or immediate feedback, so that those are some of the real added advantages of offline learning. | • It can be computer communication which could take place, so definitely there are added advantages. |
| • You have a lot of flexibility i.e. you can study on your own comfort, you can study during the night at home or at your friend’s place or anywhere that you want and definitely cost is much lower as compare to offline learning. | • You have a lot of flexibility i.e. you can study on your own comfort, you can study during the night at home or at your friend’s place or anywhere that you want and definitely cost is much lower as compare to offline learning. |

B. Providing reviews for instructors’ help:
The aim is to provide input to support lesson editors / educators / admin staff in choice-making (when student success can be strengthened) and enabling them to take the appropriate precautions. It should be remembered that this task differs from the activities of data processing and simulation, which only provide generic powerful insights from the data (articles, stats, etc.). In addition, the provision of feedback exposes entirely unique, unknown, and valuable knowledge contained in the results. A few DM techniques were used in this role, but perhaps the most common has been association rule mining. The mining of association law shows major associations between variables in huge datasets and display themselves in the face of major guidelines as per the various levels of meaning that they may have [22]. Clustering, grouping, sequential pattern analysis, modelling of dependency and simulation were used to enhance the capacity of the instructor to assess the process of learning [23].

In order to create a platform to enable the reviewer to automatically collect input from the advancement of knowledge and thus determine the efficacy of the online course, clustering, classification and association rule mining were enforced [24]. To improve the level of care and academic achievement, decision trees, bayesian models and additional statistical approaches were suggested to overcome the entrance exam [25].

C. Predicting performance among students
The predictive purpose is to estimate the unknown value of the student represented variable. Performance, ability, ranking or marking are the values generally required in education. This value is either a numerical or methodological category value. Regression analysis studies the association between a predictor variable and most significant variables. [26]. Learning is a method of classifying particular objects on the basis of quantitative knowledge about one maybe more features found in artefacts and a set of recently declassified items on the basis of learning [27]. Predicting the success of a learner is one of the elderly and the most common educational DM applications and several techniques and models (neural networks, Bayesian networks, rule-based structures, and regression and correlation analysis) have been applied.

D. Student Recommendations
The aim is to be able to give student’s feedback about their customised assignments, references to meetings, the first activity or issue that needs addressing, etc. And to be able to adapt each individual
student to the content, interfaces and learning sequences. For this role, Association rule mining, clustering and sequential pattern mining is commonly DM techniques that are used. Sequential / pattern matching mining seeks to explore the correlations among series of events, and figure out whether in the activities there is any unique order [28]. Pattern discovery mining was already evolved to customise suggestions for effective learning regard to learning style and internet usage [29], for the development of customised scenarios for learning in which the students are supported by a pattern-based method and learning patterns favoured [30], to observe the movements of the eyes (of students reading idea maps) to detect the unrelated activities coincide when focal behaviour [31], defined key operational sequences that indicate issues / achievement to help student teams recognise issues early [32], to create customised learner experiences [33], for the customisation of routes and long-term navigation effect [34], in online education framework, to suggest the most suitable potential visiting ties for a learner [35], include the definition of suggested itinerary in the SCORM norm by integrating the knowledge of teachers with the experience gained [34], to recommend examples for the student (studying things or theories) to research after that to use an integrated multimedia method [36].

Mining of association laws was used to suggest distance education tasks on a way to address or shortcuts [37], recommendation of content focused on educationally contextualised surfing habits for customised web-based learning [38], and recommendation for courseware writers on how to enhance adaptive courses [39]. In similar circumstances, clustering was designed to develop an estimation approach for potential students [40], to provide tailored content suggestions for the course based on learner capacity [41]. Other, methods of DM used are: to support automated and specialized learning, neural networks and decision trees [42], strongly advise teaching materials similar to the texts that the student has consulted, data mining and text mining [43].

E. Detecting undesirable behaviour between students

The purpose of detecting undesirable behaviour of students is to find / detect students with some type of problem or abnormal behaviours such as: wrongdoings, poor motivation, games play, overuse, gossip, drop-out, loss in academia, etc. In order to provide them with sufficient assistance in plenty of time, to expose these types of students, various data mining methods (grouping and clustering) have been used. Classification algorithms used to identify inappropriate student behaviour are decision tree neural networks, naive bayes, instance-based learning, logistic regression and SVM for learner prediction / prevention [44].

Various forms of clustering often serve to fulfil this function: Kohonen Nets to track online cheating student’s assessments [45], outlier identification to reveal a typical action of students [46], Outer detection technique to detect inappropriate learning among learners using bayesian predictive distribution [47].

Finally, for example, various data mining approaches and strategies for this assignment are: association rule mining to choose bad learners for corrective lessons [48], sending alert messages in an integrated hypermedia education system to students with unusual learning behaviour [49] and Building concept-effect relationships to diagnose learning difficulties for students [50].

F. Grouping learners

The goal is to create student groups, focused upon the personalised characteristics, interpersonal skills etc. After which the instructor/developer will use the partition of learners to create a tailored learning framework to promote productive community learning, adaptive content, etc. [3]. The Data mining methods seen in this assignment are clustering (unsupervised) and classification (supervised). Clustering or grouping is a method of grouping a set of objects in such a way that objects in the same group (called a grouping) are more similar to each other (in any sense) than objects in several other groups (clusters) [51].

Various clustering algorithms have been used to identify classes of learners with identical ability scores, such as: K means and model-based clustering [52], a broad generalised sequence-based
clustering algorithm to identify classes of learners with common learning styles based on similarities in their crossing paths and the quality of the pages that viewed [53], discriminatory settings and contextual testing roles (succeed / lose) to assist teachers in relation to information modelling [54], K-means grouping algorithm for the efficient clustering of learners with similar academic assets (task ratings, exam results and students digital classes data) [55].

Various classification algorithm has been used, such as neural network, decision tree and random forest for dividing learners into clusters (Low danger, medium danger and high risk of failure) [56], regression tree and a classification to build a model for a decision tree to explain the learning behaviour of a user to evaluate it according to various groups of cognitive styles [57].

G. Social network analysis

Instead of individual attributes or properties, Study of social networks or a systemic study aims to analyse relationships between people. A social network is considered to be an individual culture, an entity or social persons linked through social relationships like friendship, mutual connections, or sharing of information [60]. In educational settings, different strategies used to manipulate online platforms, but collective filtering is still the most important. Collaborative filtering is a way of automatically predicting a user’s desires by gathering (collaborating) taste preferences from several users [61].

Collaborative filtering has been used to inform a learner what to understand before the next step is taken. [62], to build a personal recommendation framework for learners in networks of lifelong learning [63], recommendation of appropriate relations to the active learner [64]. Online platforms research is used to identify academic collaboration groups to assist utilizing-makers in organisations to make effective decisions based upon collaboration grouping [65].

6. Conclusion

This paper is a review analysis of Educational data mining, and to date it discusses the most important work in this field. Not only by the kind of knowledge and DM approaches used, but also, and more significantly, by the kind of educational problem they overcome, and each study has been categorised. Educational data mining is becoming a growing business and one of the biggest exciting areas through which education standards can be improved digitally as well as physically, the learner’s success could be improved through making recommendations. Educational data mining is growing as illustrated in the growing entries published annually in good Journals & Conferences all over the world, and the number of specialised tasks specifically developed to apply DM techniques in academic data. So, EDM is now reaching maturity, in other words, it is no longer in its initial periods, but it is still not perfect. EDM can be useful for writers, teachers, scientists, academics, teachers by seeking reviews, etc.

References

[1] Aldowah H, Samarraie A, Mohamad F 2019 Telematics and Informatics 37, 13–49.
[2] Romero C and Ventura S 2010 IEEE Trans. Syst., Man, Cybern. 40, 601-618.
[3] Castro F, Vellido A, Nebot A, Mugica F 2007 Studies in Computational Intelligence, 62, 183-221.
[4] Mostow J, Beck J 2006 Journal Natural Language Engineering, 12, 195-208.
[5] Koedinger K, Cunningham K, Skogsholm A, Leber B 2008 1st International Conference on Educational Data Mining, Montreal, 157-166.
[6] Guruler H, Istanbullu A, and Karahasan M 2010 Computer Education, 55, 247–254.
[7] Ryan S, Baker K, Kalina Y 2009, Journal of Educational Data Mining.
[8] Alejandro P 2014, Expert Syst. Appl., 41, 1432-1462.
[9] Baker R 2010, International Encyclopedia of Education 3rd edition. Oxford, UK: Elsevier.
[10] Romero C, Ventura S, De Bra P (2004). Knowledge discovery with genetic programming for providing feedback to courseware author. User Modeling and User-Adapted Interaction: The Journal of Personalization Research vol.14, no. 5, 425–464.
[11] Raghavan, S.N.R. (2005). Data Mining in E-commerce: A Survey Sadhana Journal, vol. 30, 2&3, 275–289.
[12] Romero C, Ventura S (2007). Educational Data Mining: a Survey from 1995 to 2005. Expert Systems with Applications, vol. 1, 33, 135-146.
Kotsiantis, S., Pierrakeas, N., Nagata, R., Takeda, K., Suda, K., Kakegawa, J., Morihiro, K. (2009). Edu Guo, Q., Zhang, M. (2009). Implement web learning environment based on data mining. In Lu, F., Li, X., Liu, Q., Yang, Z., Tan, G., He, T. (2007). Research on Personalized E-Learning System Using Fuzzy Set Based Clustering Algorithm. In International Conference on Computational Science, Beijing, China, 587-590.

24. Guo, Z. (2009). Implement web learning environment based on data mining. In Knowledge-Based Systems Journal, 22, 439-442.

25. Nagata, R., Takeda, K., Suda, K., Kakegawa, J., Morihiro, K. (2009). Edu-mining for book recommendation for pupils. In International Conference on Educational Data Mining, Cordoba, Spain, 91-100.

26. Kotsiantis, S., Pierrakeas, C., Pantelis, P. (2003). Preventing student dropout in distance learning systems using machine learning techniques. In International Conference on Knowledge-Based Intelligent Information &
[45] Burlak, G., Muñoz, J., Ochoa, A., Hernández, J.A. (2006). Detecting Cheats In Online Student Assessments Using Data Mining. In *International Conference on Data Mining*, Las Vegas, 204-210.

[46] Vee, M.N., Meyer, b., Mannock, K.L. (2006). Understanding novice errors and error paths in Object-oriented programming through log analysis. In *Workshop on Educational Data Mining*, Taiwan, 13-20.

[47] Nagaoka, K. (2002). Learning log database and data mining system for e-learning – on line statistical outlier detection of irregular learning processes. In *International Conference on Advanced Learning Technologies*, Tatarstan, Russia 436-438.

[48] Ma, Y., Liu, B., Wong, C., Yu, P. Lee, S. (2000). Targeting the right students using data mining. In *KDD ’00: Proceedings of the sixth ACM SIGKDD International Conference on Knowledge discovery and data mining*, 457–464.

[49] Jong, B.S., Chan, T.Y., Wu, Y.L. (2007). Learning log explorer in elearning diagnosis. In *IEEE Transactions on Education Journal*, 50, 216-228.

[50] Hwang, G.J. (2005). A data mining approach to diagnosing student learning problems in science courses. In *Journal of Distance Education Technologies*, 3, 35-50.

[51] Romesburg, H.C. (2004). Cluster Analysis for Researchers. Krieger Pub.

[52] Ayers, E., Nugent, R., Dean, N. (2009). A Comparison of Student Skill Knowledge Estimates. In *International Conference On Educational Data Mining*, Cordoba, Spain, 1-10.

[53] Tang, T. Y., Mccalla, G. (2002). Student modeling for a web-based learning environment: a data mining approach. In *Conference on Artificial intelligence*, Edmonton, Canada, 967-968.

[54] Gaudioso, E., Montero, M., Talavera, L., Hernandez-del-olmo, F. (2009). Supporting teachers in collaborative student modeling: a framework and an implementation. In *Expert System with Applications*, 36, 2260-2265.

[55] Chen, c., Chen, M., Li, Y. (2007). Mining key formative assessment rules based on learner port files for web-based learning systems. In *IEEE International Conference on Advanced Learning Technologies*, Japan, 1-5.

[56] Superby, J.F., Vandamme, J.P., Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *International conference on intelligent tutoring systems*, Educational Data Mining Workshop, Taiwan, 1-8.

[57] Lee, M.W., Chen, S.Y., Liu, X. (2007). Mining learners’ behaviour in accessing web-based interface. In *International Conference Edutainment*, Hong Kong, China, 226-346.

[58] Merceran, A., Yacef, K. (2005). Educational Data Mining: a Case Study. In *International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands, 1-8.

[59] Hanna, M. (2004). Data mining in the e-learning domain. In *Campus- Wide Information Systems, Volume 21*, Number 1, 29-34.

[60] Freeman, L. (2006). The Development of Social Network Analysis. Empirical Press.

[61] Herlocker, J., Konstan, J., Tervin, L.G., Riedl, J. (2004). Evaluating collaborative filtering recommender systems, *ACM Transactions on Information Systems Journal*, 22, 1, 5-53.

[62] Zhu, F, Ip, H., Fok, A., Cao, J. (2007). Pe RES: A personalized recommendation education system based on multi-agents & SCORM. In *International Conference on Web-based Learning*, Edi burgh, 31-42.

[63] Drachsler, H., Hummel, H. G., Koper, R. (2008). Personal recommender systems for learners in lifelong learning networks: the requirements, techniques and model. In *International Journal of Learning Technology*, 3, 4, 404-423.

[64] Khribi, M. K., Jenni, M., Nasraoui, O. (2008). Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. In *IEEE International Conference on Advanced Learning Technologies*, Washington, DC, 241-245.

[65] Nankani, E., Simoff, S. Denize, S., Young, L. (2009). Supporting Strategic Decision Making in an Enterprise University through Detecting Patterns of Academic Collaboration. In *International United Information Systems Conference*, Sydney, Australia, 496-507.

[66] Baker RS, Yacef K (2009) The state of educational data mining in 2009: a review and future visions. JEDM-J, *Educational Data Mining* 1(1):2017