# Educational Data Mining –Applications and Techniques

B.Namratha
*Assistant Professor, Department of Information Technology,*
*Anurag Group of Institutions,RR Dist., Hyderabad, Telangana, India*

Niteesha Sharma
*Assistant Professor, Department of Information Technology,*
*Anurag Group of Institutions,RR Dist., Hyderabad, Telangana, India*

**Abstract - Data mining is referred to the process of extracting hidden and useful information in large data repositories. Knowledge Discovery and Data Mining (KDD) is a multidisciplinary area focusing upon methodologies for extracting useful knowledge from data and there are several useful KDD tools to extracting the knowledge. This knowledge can be used to increase the quality of education. Educational Data Mining is concerned with developing new methods to discover knowledge from educational/academic database and can be used for decision making in educational/academic systems. This paper discusses about what is educational data mining, its broad application areas, benefits of educational data mining, challenges and barriers to successful application of educational data mining and the new practices that have to be adopted in order to successfully employ educational data mining and learning analytics for improving teaching and learning.**

**Keywords: Data Mining, Educational Data Mining, Knowledge Discovery in Database (KDD).**

## I. INTRODUCTION

Educational Data Mining (EDM) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational settings (e.g., universities and intelligent tutoring systems). EDM refers to techniques and tools designed for automatically extracting meaning from large repositories of data generated by peoples learning activities in educational settings At a high level, the field seeks to develop and improve methods for exploring this data, which often has multiple levels of meaningful hierarchy, in order to discover new insights about how people learn in the context of such settings In doing so, EDM has contributed to theories of learning investigated by researchers in educational psychology and the learning sciences. The field is closely tied to that of learning analytics, and the two have been compared and contrasted. Quite often, this data is extensive, fine-grained, and precise. The main objective of applying Data Mining to educational data is to analyze educational Data contents, models, to summarize/analyze the learner's discussions, etc. Education Data Mining concentrates on the computing process models which focus on Education context.

In educational system, a student's performance is determined by the term work, attendance and end semester examination. The term work is carried out by the teacher based upon student's performance in educational activities such as class test, assignments, attendance. The end semester examination is one that is scored by the student in semester examination. Student has to get minimum marks to pass a semester in internal as well as end semester examination.

*Calders & Pechenizkiy (2011) associate basic EDM tasks to traditional data mining problems, i.e.:*

| Classic DM Problems | Educational Example |
|---|---|
| Classification | Categorizing and profiling students, determining their learning styles and preferences. |
| Predictive Modeling | Inducing models that can predict whether (and when) a student will pass a course or not or will eventually graduate or drop out. |
| Clustering | Grouping Similar students (based on behavior, performance, etc) or grouping similar courses, assignments etc together, exploring collaborative learning patterns. |
| Bi-Clustering | Finding which questions (tasks, courses etc) are difficult/easy for which students. |
| Frequent Pattern Mining | Finding (elective) courses often taken together or popular paths in study programs or actions in LMS. |
| Emerging Pattern Mining | Finding patterns that capture significant differences in behavior of students who graduated vs. those students who did not or that explain the changes in behavior of student generations over different years. |
| Collaborative filtering and recommendations | Recommending suitable learning objects, based on the analysis of the performance of other learners, recommending remedial classes for the students. |
| Visual Analytics | Facilitating reasoning about the educational processes or learning results via interactive data/model visualization, e.g. Visualizing collaborations of students. |

## II. APPLICATIONS OF DATA MINING IN HIGHER EDUCATION

List of the primary applications of EDM is provided by Cristobal Romero and Sebastian Ventura. In their taxonomy, the areas of EDM application are:

- Analysis and visualization of data
- Providing feedback for supporting instructors
- Recommendations for students
- Predicting student performance
- Student modeling
- Detecting undesirable student behaviors
- Grouping students
- Social network analysis
- Developing concept maps
- Constructing courseware
- Planning and scheduling

There are many application areas of data mining like customer analytics, Agriculture, banking, Security Applications, Educational data mining, Mass surveillance, Privacy preserving etc. The main concerned area is about data mining applications in educational systems. Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.
A key area of EDM is mining student's performance. Another key area is mining enrollment data. Key uses of EDM include predicting student performance and studying learning in order to recommend improvements to current educational practice. EDM can be considered one of the learning sciences, as well as an area of data mining.
The main applications of EDM are listed as follows:

*A. Analysis and Visualization of Data*

It is used to highlight useful information and support decision making. In the educational environment, for example, it can help educators and course administrators to analyze the students' course activities and usage information to get a general view of a student's learning. Statistics and visualization information are the two main techniques that have been most widely used for this task. Statistics is a mathematical science concerning the collection, analysis, interpretation or explanation, and presentation of data. It is relatively easy to get basic descriptive statistics from statistical software, such as SPSS. Statistical analysis of educational data (logs
files/databases) can tell us things such as where students enter and exit, the most popular pages students browse, number of downloads of e-learning resources, number of different pages browsed and total time for browsing different pages. It also provides knowledge about usage summaries and reports on weekly and monthly user trends, amount of material students might go through and the order in which students study topics, patterns of studying activity, timing and sequencing of events, and the content analysis of students notes and summaries. Statistical analysis is also very useful to obtain reports assessing how many minutes student worked, number of problems here solved and his correct percentage along with our prediction about his score and performance level.

Visualization uses graphic techniques to help people to understand and analyze data. There are several studies oriented toward visualizing different educational data such as patterns of annual, seasonal, daily and hourly user behavior on online forums. Some of such investigations are statistical graphs to analyze assignments complement, questions admitted, exam score, student tracking data to analyze student's attendance, results on assignments and quizzes, weekly information regarding students and group's activities.

*B. Predicting Student Performance*

In this case, we estimate the unknown value of a variable that describes the student. In education, the values normally predicted are student's performance, their knowledge, score, or marks. This value can be numerical/continuous (regression task) or categorical/discrete (classification task). Regression analysis is used to find relation between a dependent variable and one or more independent variables. Classification is used to group individual items based upon quantitative characteristics inherent in the items or on training set of previously labeled items. Prediction of a student's performance is the most popular applications of DM in education. Different techniques and models are applied like neural networks, Bayesian networks, rule based systems, regression, and correlation analysis to analyze educational data. This analysis helps us to predict student's performance i.e. to predict about his success in a course and to predict about his final grade based on features extracted from logged data. Different types of rule-based systems have been applied to predict student's performance (mark prediction) in an elearning environment (using fuzzy-association rules).Several regression techniques are used to predict student's marks like linear regression for predicting student's academic performance, stepwise linear regression for predicting time to be spent on a learning page, multiple linear regression for identifying variables that could predict success in colleges courses and for predicting exam results in distance education courses.

*C. Grouping Students*

In this case groups of students are created according to their customized features, personal characteristics, etc.

These clusters/groups of students can be used by the instructor/developer to build a personalized learning system which can promote effective group learning. The DM techniques used in this task are classification and clustering. Different clustering algorithms that are used to group students are hierarchical agglomerative clustering, *K*-means and model-based clustering. A clustering algorithm is based on large generalized sequences which help to find groups of students with similar learning characteristics like hierarchical clustering algorithm which
are used in intelligent e-learning systems to group students according to their individual learning style preferences.

*D. Enrollment Management*

This term is frequently used in higher education to describe well-planned strategies and tactics to shape the enrollment of an institution and meet established goals. Enrollment management is an organizational concept and a systematic set of activities designed to enable educational institutions to exert more influence over their student enrollments. Such practices often include marketing, admission policies, retention programs, and financial aid awarding. Strategies and tactics are informed by collection, analysis, and use of data to project successful outcomes. Activities that produce measurable improvements in yields are continued and/or expanded,

While those activities that do not are discontinued or restructured. Competitive efforts to recruit students are a common emphasis of enrollment managers. The numbers of universities and colleges instituting offices of

"enrollment management" have increased in recent years. These offices serve to provide direction and coordination of efforts of multiple offices such as admissions, financial aid, registration, and other student services. Often these offices are part of an enrollment management division. Some of the typical aims of enrollment management include

- Improving yields at inquiry, application, and enrollment stages.
- Increasing net revenue, usually by improving the proportion of entering students capable of paying most or all of unsubsidized tuition.
- Increasing demographic diversity
- Improving retention rates
- Increasing applicant pools

## III. TECHNIQUES OF EDUCATIONAL DATA MINING

### *A. Clustering Algorithm*

Clustering is a division of data into groups of similar objects. Clustering plays an outstanding role in data mining applications such as information retrieval and text mining ,scientific data exploration, web analysis, spatial database applications, medical diagnostics, marketing and many more..

Data Clustering is unsupervised and statistical data analysis technique. It is used to classify the same data into a homogeneous group of primary school students it is used to operate on a large data-set to discover hidden pattern and relationship helps to make decision quickly and efficiently. Cluster analysis is used to break down a large set of data into subsets called clusters. Each cluster is a collection of data objects that are similar to one another. They are placed within the same cluster but are dissimilar to objects in other clusters. Following algorithms are used in education mining in Clustering.

### *B. K-Mean Clustering Algorithm*

The K-means is one of the best clustering algorithms in data mining. K-Means is a non-hierarchical clustering method that seeks to partition the data into the form of one or more clusters .This method partitions the data into clusters so that the data having the same characteristics are grouped into one cluster and the data that have different characteristics grouped into another cluster.

K-Means Clustering (KMC) proposes to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. Exactly *k* different clusters have been produced by this method with greatest possible characteristic. Initially best number of clusters *k* leading to the greatest separation (distance) is not known and must be computed from the data. K-Meansclustering's objective is to minimize the squared error function or total intra-cluster variance.

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'ci'* represents the number of data points in *ith* cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

### C. *Classification*

Classification is the form of data analysis that extracts models describing important data classes. This approach frequently uses decision tree classification algorithms. The data classification process includes learning and classification. In learning method, the training data sets are analyzed by the classification algorithm. In classification

test data sets are used to find the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples.

### *D. ID3 Algorithm*

Terminologies used in ID3 Algorithm:

- Establish Classification Attribute
- Compute Classification Entropy.
- Calculate Information Gain using classification attribute.
- Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).
- Remove Node Attribute, creating reduced table.

### *E. Dimensionality reduction techniques*

The dimensionality reduction is the utmost vital method to eliminate redundant attributes and noise which can be further classified into feature extraction and assortment method. The student clusters are generated from the well-known COBWEB algorithm. The student clusters are formed based on the credits obtained from the given semester. As stated before, the grouping utility is defined as the weighted distance utility of the attribute (i.e. credits).To mine the students' performance data, the data mining classification techniques such as – Decision tree-Random Tree and J48 classification models were built with 10 cross validation fold using WEKA.

## IV. CONCLUSION

This paper discusses about educational data mining, its applications and techniques to be used in educational data mining. The application of data mining methods in the educational sector is an interesting phenomenon. Data mining techniques in educational organizations help us to learn student performance, student behavior, designing course curriculum and to motivate students on various parameters.

## REFERENCES

[1] Jiawei Han and Micheline Kamber, Data Mining- Concepts and Techniques: Elsevier Publishers, 2006.
[2] Crist´obal Romero and Sebasti´an Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man, and Cybernetics—Part c: Applications and Reviews, vol. 40, no. 6, 2010, pp. 601-618.
[3] Richard A. Huebner, Norwich University, "A Survey on Educational Data Mining", Research in Higher Education Journal. (www.aabri.com).
[4] B.K. Bharadwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance", International Journal of Advance Computer Science and Applications, Vol. 2, No. 6, 2011.
[5] Ajay Kumar Pal1, Saurabh Pal, "Evaluation of Teacher's Performance: A Data Mining Approach", IJCSMC, Vol. 2, Issue. 12, December 2013.
[6] Aggarwal, C. Charu and Yu, S. Philip. "Data Mining Techniques for Associations, Clustering and Classification." in Zhong, Ning and Zhou, Lizhu (Eds.) methodologies for knowledge discovery and data mining, third pacific Asia Conference, PAKDD, Beijing,China,April26-28 1999 proceedings, Springer, New York
[7] Dr. Mohd Maqsood Ali, Jazan University, "Role Of Data Mining In Education Sector", IJCSMC, Vol. 2, Issue. 4, April 2013.
[8] Baker, R., & Yacef, K. (2009). The State of Educational Data mining in 2009: A Review Future Visions. *Journal of Educational Data Mining.*