

Data Mining in Education with Virtual Learning Environment Data

Jasti Sri Radhe Shyam

Computer Science and Technology Department
HMR Institute of Technology and Management
New Delhi, India

Shelly Goyal

Computer Science and Technology Department
HMR Institute of Technology and Management
New Delhi, India

Ashish Kumar

Computer Science and Technology Department
HMR Institute of Technology and Management
New Delhi, India

Kamal Kumar Arya

Computer Science and Technology Department
HMR Institute of Technology and Management
New Delhi, India

Abstract—There is an increase in e-learning using VLE (Virtual learning environment) over years and there is also increase in understanding the teacher, student and also management relationships in a institution. This study is done by data mining and when it is used for educational purposes it is known as EDM (Educational Data Mining). The data acquired from the VLE processed by Data Mining Techniques give us a certain patterns and characteristics of individual and their online interaction which can be used for further enhancing education system in a institution or education society. This paper aim is to make familiar with data mining concept, data mining techniques used in EDM and usage of these techniques with VLE data.

Index Terms— VLE, EDM, Data Mining

1. INTRODUCTION

Our world got many changes over the years in implementing and evaluating the education system for improvements in education paradigm.

In this era, there is a drastic increase in population, diversity, globalisation, education system, and in science and technology advancements. Due to this increase, there is a problem of gathering, storing and analysing the large pool of data, so technology also got advanced and some are invented to decrease this problem. The world is changing at faster pace hence the education system and the syllabus should also adapt to the change. To do so we have to understand the current trends, ideas and methodologies to adapt the changes accordingly, this can be done by Data Mining.

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [1]. Data Mining is used in many fields from atom related data to space exploration data and many other fields where we want to find patterns that affect our current knowledge about a certain topic we are interested in.

Over the years there is an increase in interest in the scientific study of education system using data mining, and this is termed as educational data mining (EDM). Educational data mining

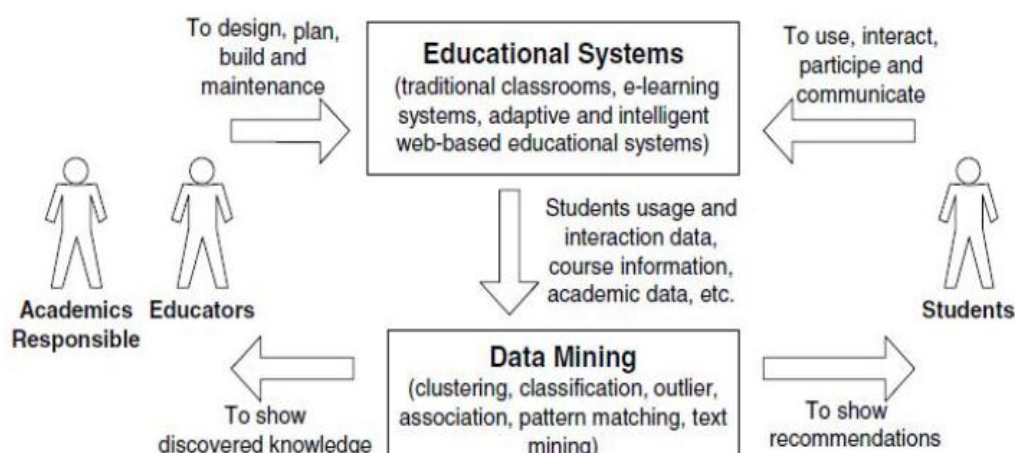


Figure: The cycle of applying data mining in educational system, (Source: A Survey and Future Vision of Data mining in Educational Field [18])

(also referred to as “EDM”) is defined as the area of scientific inquiry centred on the development of methods for making discoveries within the unique kinds of data that come from educational settings and using those methods to better understand students and the settings which they learn in [1].

There is a use of VLE (Virtual Learning Environment) in many educational and corporate organisations. VLE examples are Moodle (Open Source), Blackboard (proprietary), etc. In the era of digitalisation, VLE is becoming an integral part of many educational institutions and we can use the data generated by VLE during student, teacher and management interactions with different services in the VLE system. The data generated can be used for data mining for determining and obtaining different relations and patterns in data. Next section gives common Educational Data Mining techniques and then Case studies for an idea how Data Mining can be used with VLE.

2. POPULAR EDM METHODS

Before According to Romero, Ventura and Baker, the methods of EDM are categorised into:

- Prediction
- Clustering
- Relationship Mining
- Text Mining
- Social Network Analysis

The first three categories are universal categories of data mining and are always used for all forms of data mining. The rest of two other categories are usually used for EDM only.

2.1 Prediction

The goal is to develop a model which can infer a single aspect of the data (predicted variables). Labels are required for the output variable, where a label represents some information about the output variables value. It has two key uses within EDM. In the first type, prediction methods can be used to study what are the features of a model. It is generally used to predict student educational outcomes. In the second type, prediction methods are used to predict what the output value would be in contexts where it is not desirable to directly obtain a label for that construct.

Prediction can be classified into three types:

- Classification
- Regression
- Density Estimation

Classification is a task in data mining. Classification predication encompasses two levels: Classifier construction and the usage of the classifier constructed. The entire process begins with the collection of evidence acquired from various data sources or warehouses. In the ideal situation, the data should be of low-dimensionality, independent and discriminative so that its value similar to characteristics in the same class but very in features from different classes. The building of a classification process model can be broken into four component technique choice, data pre-processing, training, and testing or evaluation. Some popular classification methods include logistic regression, support vector machines and decision trees.

In regression, the predicted variable is a continuous variable. Some popular regression methods within EDM include linear regression, neural networks and support vector

machines regression. Neural Network (NN) is made up of structure or a network of numerous interconnected units (Artificial Neurons). Each of these units consists of input/output characteristics that implemented a local computation or function. The function could be a computation of weighted sums of inputs which produce an output if it exceeds a given threshold. The output could serve as an input to other neurones in the network. This process iterates until a final output is produced.

In density estimation, the predicted variable is probability density function. It also includes Gaussian function.

2.2 Clustering

It is a process of grouping objects into classes of similar objects. Students typically annotate texts while reading a book by highlighting the context of interest or by underlying it or by writing comments in the side margins. This activity is called annotation [8]. Researchers have applied statistical clustering method like k-means clustering and Hierarchical Clustering to student annotations. And they proved that by using their clustering methods, the creation of students with similar learning style cluster is improved and is faster.

Learning Management Style (LMS) have become an integral part of educational institutions for teaching and learning. A typical LMS logs most of the user activities like course attempted, modules read, practice exam attempted, exam score, chat logs of student – student interaction or student – teacher interaction.

2.3 Relationship Mining

In Relationship Mining, the goal is to discover relationships between variables, in a dataset with a large number of variables. This may take the form of attempting to find out which variables are most strongly related or associated with a single variable of particular interest. Broadly relationship mining is classified into four types:

- Association Rule mining
- Correlation mining
- Sequential pattern mining
- Casual data mining

Association rule mining discovers relationships among attributes in the data set, producing if-then statements concerning attribute-values [6]. Association rule mining is one of the important technique which aims at extracting, interesting correlations, frequent patterns. Association rule mining has been applied to EDM for finding students mistakes often occurring together while solving exercises.

In correlation mining, the goal is to find (positive or negative) linear correlations between variables. It defines how two data values or attributes related to each other.

Sequential pattern mining attempts to find inter-session patterns such as the presence of a set of items followed by another item in a time-ordered set of sessions.

Wang et al. [5] propose a four-phase learning portfolio mining approach, which uses sequential pattern mining, to extract learning features to create a decision tree which is used to predict which group a learner belongs to.

Casual data mining attempts to find whether one event was the cause of another event, either by analysing the covariance of two events or by using information about how one of the events was staggered.

2.4 Text Mining

Text mining can be viewed as an extension of data mining to text data and it is much related to web content mining. Text mining can work with semi-structured or unstructured data sets such as text documents, HTML files, Emails etc. Ueno uses data mining and text mining technologies for collaborative learning and discussion brown with evaluation between peers in an ILMS. Chen et al. [2] propose to automatically construct e-textbooks via web content mining. The specific application of text mining in e-learning can be used for evaluating the progress of the thread discussion to see what the contribution to the topic is. Identifying the main blocks of multimedia presentations and retrieve their internal properties[11].

2.5 Social Network Analysis (SNA)

It is a field of study attempting to understand and measure relationships between entities in networked information. The SNA techniques and data mining techniques for information networks can be used to examine and assess online interactions [3]. It uses the connection among units as data which relate them to one another data units in a network. Rallo et al. [4] propose to use that data mining and social networks to interpret and analyse the structure and contents of online educational communities.

3. CASE STUDIES

Before Here we studied two cases in the first case authors tried to extract the data manually from VLE and analyse the data extracted to find student behaviour patterns. And other case tries to simplify the data by creating a plugin to extract the data in XML format which most popular intermediary data representation format and analyse on the extracted data. In both the cases, same VLE is used based on Moodle [10].Here are two cases:

3.1 Student Behaviour Patterns in a VLE

In this case, authors tried to identify student behaviour patterns that are obtained from their interactions on a VLE via VLE at the “Utniversidad Técnica Particular de Loja” by Priscile Valdiviezo, Ruth Reátegui, Marcia Sarango [12]. They used Clustering techniques to classify certain indicators and to obtain groups of students with similar characteristics [12]. Their research included 388 students from different courses. The data that was analysed was the log registers generated by VLE during the student's interaction while using different content sharing services available in Virtual Platform. Then they tried to select the entities that held most useful information about students actions performed and which represented student's interaction with the system on the virtual platform.

Then they established two key indicators:

I. Course Participation

Measuring the contribution by students interaction made by students while enrolled in a particular course

II. Usage of online tools

To address the student's action and usage of online forums, instant messaging, chats, online resources, usage of twitter, etc.

Then two indicators were measured using three criteria:

- Permanent (P), referring to a high level of participation and interaction with the tools.
- Moderate (M), referring to a medium level which includes both interaction and usage of online tools.

- Low (E), referring to minimum values (low access and minor usage of tools) during the course.

Then they used WEKA tool for data mining processing with Cluster K-means Algorithm. With the tool, they created three groups on basis of cluster generated which was based on the participation of students in VLE. They name clusters as Group1 – with higher participation, Group2 – with a medium level of participation, Group3 – with a low level of participation.

The results obtained indicated that the greatest level of student interaction was in the forum, followed by quizzes, online tasks, instant messaging, the usage of online resources and twitter. They also observed a group of students who were at risk of not competing in a particular course or subject. Likewise, some students might fall behind in their academic studies. These results make more informed decisions about their teaching practices on VLE to professors and tutors.

3.2 Moodle data retrieval for Education Data Mining

In this case, author (Felermino) tries to add or create a new service into existing (LMS) Moodle Core Services for Student's usage data retrieval. The author wants to incorporate this new service to ease the pre- processing and provide a more generic mechanism for data portability.

There are already two tools at the time author wrote the paper to facilitate data extraction in Moodle, first MMT (Moodle Mining Tool) [17] and ADE (Automatic data extraction). Both tools adapt accordingly to a particular data mining tool and implemented over a framework [9]. Moodle with data mining frameworks usually sets boundaries in terms of data portability [9]. According to the author data extracted using ADE in MMT requires further steps of data transformation due to the fact that dataset format depends heavily upon the data mining tool. So, the author uses XML format to set out the limitations as it is used throughout the web for intermediary data representation. This format will provide accessibility to a wide range of platforms.

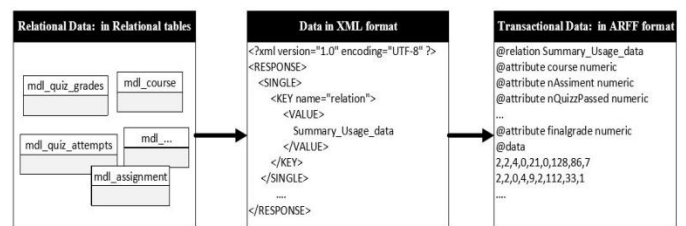


Figure: Data transformation flow (Source: Moodle Data Retrieval for Educational Data Mining [9])

This service created by Felermino will transform Moodle (LMS) relational data into a dataset by merging several tables into one table, and then return the table as Key-Value pair XML that contains two parts, known as header and data. Header - contains the name of the relation, a list of the attributes and attribute types and Data – contains the information. Then JABX (Java Architecture for XML Binding) mechanism will transform the returned XML into ARFF representation for data mining.

4. SOME OTHER RELATED WORKS

Before There have been many studies and work going on EDM(Educational Data Mining) over the years and some of them are discussed here to get a glimpse of EDM for future work.

The activities and results that should be considered when analysing the student's interaction with the system: tasks performed, the order and time of activities performed, the percentage of exercises performed correctly [13]. Brusilovsky and Millan focused on 5 features with respect to the user as an individual – the user's knowledge, goals, interest, background and individual traits.

Pertrushyna et al.,[14] analysed student's interaction in the forum for patterns of student behaviour and their role in the community who are learning English. Blikstein [15] describes a method to evaluate, analyse, and visualise students who learn computer programming. He uses snapshots of code in programs source code and different quantitative techniques to extract information about student behaviour and classifies this information in terms of their experience in programming. Anaya and Boticario [16], data mining techniques were applied to statistical indicators of student interactions in the forums of a VLE, to obtain information about the group collaboration. And there are many other examples data mining on the data generated by VLE.

5. CONCLUSION

This paper presented some commonly used techniques in EDM and Case Studies that showed how an EDM is done for a particular dataset and other showed how to create a custom piece of software for easily understandable and inter-portable dataset for EDM from VLE. And after that, some other examples were discussed for an idea about other data mining procedures and outcomes. The main aim of this paper is to give an idea to the reader about the data mining in the educational field using a VLE platform. As the number of students increasing there is a problem of assessing each one of them is difficult with a traditional method so to keep up with students we can use a VLE for gathering the data and analyse the data with the help of data mining.

REFERENCES

- [1] Baker, R.S.J.d. (in press) "Data Mining for Education." In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.
- [2] Chen, J., Li, Q., Wang, L., & Jia, W., "Automatically generating a textbook on the web" In International conference on advances in web-based learning, Beijing, China, 2004, (pp. 35–42).
- [3] Scott, J. Social network analysis: A handbook (2nd ed.). Newberry Park, CA: Sage.2000.
- [4] Rallo, R. Gisbert, M., & Salinas, J., "Using data mining and social networks to analyze the structure and content of educative online communities" In International conference on multimedia and ICTs in education, Caceres, Spain, 2005, (pp. 1–10).
- [5] Wang, W., Weng, J., Su, J., & Tseng, S., "Learning portfolio analysis and mining in SCORM compliant environment" In ASEE/ IEEE frontiers in education conference, 2004, (pp. 17–24).
- [6] Agarwal, R., Imielinski, T., & Swami, A., "Mining association rules between sets of items in large databases" In Proceedings of the ACM SIGMOD international conference on management of data, Washington DC, USA, 1993, (pp. 1–22).
- [7] Agarwal, R., & Srikant, R., "Mining sequential patterns" In Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan, 2005, (pp. 3–14).
- [8] Romero, C., Ventura, S., Espejo, P.G. and Hervas, C., "Data Mining Algorithms to Classify Students" In Proceedings of the 1st International Conference on Educational Data Mining, 2008, (pp. 8-17).
- [9] Felermino M.D.A.Ali, "Moodle Data Retrieval for Educational Data Mining". In International Journal of Scientific Engineering and Technology, 2015, (pp: 523-525).
- [10] Moodle [Internet]., 2017, Available from: <https://moodle.org> .
- [11] Bari, M., & Benzater, B., "Retrieving data from pdf interactive multimedia productions," In International conference on human system learning: Who is in control? 2005, (pp. 321–330).
- [12] Priscile Valdiviezo, Ruth Reátegui, Marcia Sarango, "Student Behavior Patterns in a Virtual Learning Environment", In Eleventh LACCEI Latin American and Caribbean Conference for Engineering and Technology, 2013.
- [13] Peredes, P., Rodríguez, P. A. (2004). "A mixed approach to modelling learning styles in adaptive educational hypermedia". Proceedings IASTED Conference on Web-Based Education (WBE 2004), (pp. 16-18). February 2004.
- [14] Petrushyna, Z., Kravcik, M., Klamma, R. (2011). "Learning analytics for communities of lifelong learners: a forum case". Proceedings In: 11th IEEE International Conference on Advanced Learning Technologies, (pp. 609-610).
- [15] Blikstein, P. (2011). "Using learning analytics to assess students' behavior in open-ended programming tasks". Proceedings I Learning Analytics Knowledge Conference (LAK 2011). (pp. 110–116), Feb. 2011.
- [16] Anaya, A. R., Boticario, J. G. (2009). "Clustering Learners according to their Collaboration". Proceedings Computer Supported Cooperative Work in Design (CSCWD 2009), (pp. 540–545), April 2009.
- [17] Taylor P, Jovanovic M, Vukicevic M, Milovanovic M, Minovic M. Using data mining on student behavior and cognitive style data for improving e-learning systems : a case study. Int J Comput Intell Syst. 2012;(May 2012): (pp.37–41).
- [18] Barahate Sachin R., Shelake Vijay M., "A Survey and Future Vision of Data mining in Educational Field", in Second International Conference on Advanced Computing & Communication Technologies (2012).