Education Data Mining Using R Programming

Dhiraj Upadhyaya
Amity University
Noida, Uttar Pradesh, India
dupadhyaya@amity.edu

Anchal Garg
Amity University,
Noida, Uttar Pradesh, India
agarg@amity.edu

ABSTRACT

Data mining is the process of finding interesting patterns in data. With increased use digital technology, large data has got generated from the activities which take place in educational domain. Education Data Mining (EDM) is the application of data mining techniques on this educational data. These techniques include prediction, classification, association rule analysis. R is an open source programming language and software environment for statistical analysis, graphical representation and reporting. This paper brings out how R Programming can be used perform EDM.

Keywords: Data mining, education data mining, learning analytics, learning analytics, academic analytics, r programming, classification, prediction

INTRODUCTION

Education is the primary activity which every human being undergoes for their development. Education consists of various activities like attendance, assessment, group activities etc. Every teacher and institution want to measure the effectiveness of teaching and institution performance. Data is the important source for such measuring which can be processed for further techniques.

Collection and analysis of data about learning is a trend that is growing at all levels of education. (Lodge & Corrin, 2017). This data when analysed can provide greater insights about student learning. With growth in digital technology, there is exponential growth of education data which has become the biggest challenge in educational institutions. (Cheng, 2017). Data in plain form is of no value unless some techniques are applied on it. Data mining is broad area that includes using different techniques and algorithms for identifying patterns. (Nithya, Umamaheswari, & Umadevi, 2016). Data mining on Educational data can transform the invisible, unnoticed even useless data become visible and understandable (Bienkowski, Feng, & Means, 2012).

This paper is to synthesize and share how R programming can be used for various EDM techniques. It explains EDM and then brings out ways in which R programming is applied for data mining to educational data

EDUCATIONAL DATA MINING

Educational data mining is the application of data mining methods and tools to education related data typically collected through the use of an e-learning platforms. (Charitopoulos & Rangoussi, 2017). This has become an emerging discipline, however its methods is often different from those methods from the broader data mining literature. (Cheng, 2017) It is a multi-disciplinary field which includes data mining, learning theory, data visualisation, machine learning and psychometrics. (Roy & Garg, 2018). Due to this uniqueness it can be used to solve educational related issues.

International Journal of Education Data Mining (JEDM) is the one of the first journal dedicated to EDM research. An yearly international EDM conference by these researchers since 2008. (Baker & Yacef, 2009). Goals of EDM (Baker & Yacef, 2009) are predicting students future learning behavior, discovering or

ISBN: 978-93-5351-521-8

improving domain models, studying the effects of educational support and advancing scientific knowledge about learning and learners. EDM can help both students and management for improving the quality of education. Student data mining is the mining of student data or data related to the students for ex. courses assignments, marks, student background etc. (Kumar, 2011). What if scenario with student data can help in decision making process. Mining on student data can provide better perspective throughout the education processes and also analyze information related to programs, courses and course assignments. EDM can also highlight educational processes that need improvement and also assist in design of educational content. EDM can guide optimal utilization of educational resources thereby helping the management.
Key components of EDM identified for achieving Educational objectives are stakeholders of education, DM methods and techniques, Education data and Educational tasks and outcomes. (Jindal & Borah, 2013). In this paper authors explain all these components.

Learning Analytics (LA) is a related area which explores data during teaching and learning process. It can be defined as collection, analysis, and use of data associated with student learning (Roy & Garg, 2017). Though EDM and LA are similar, they attempt to answer different questions which have been summarized in figure 1.

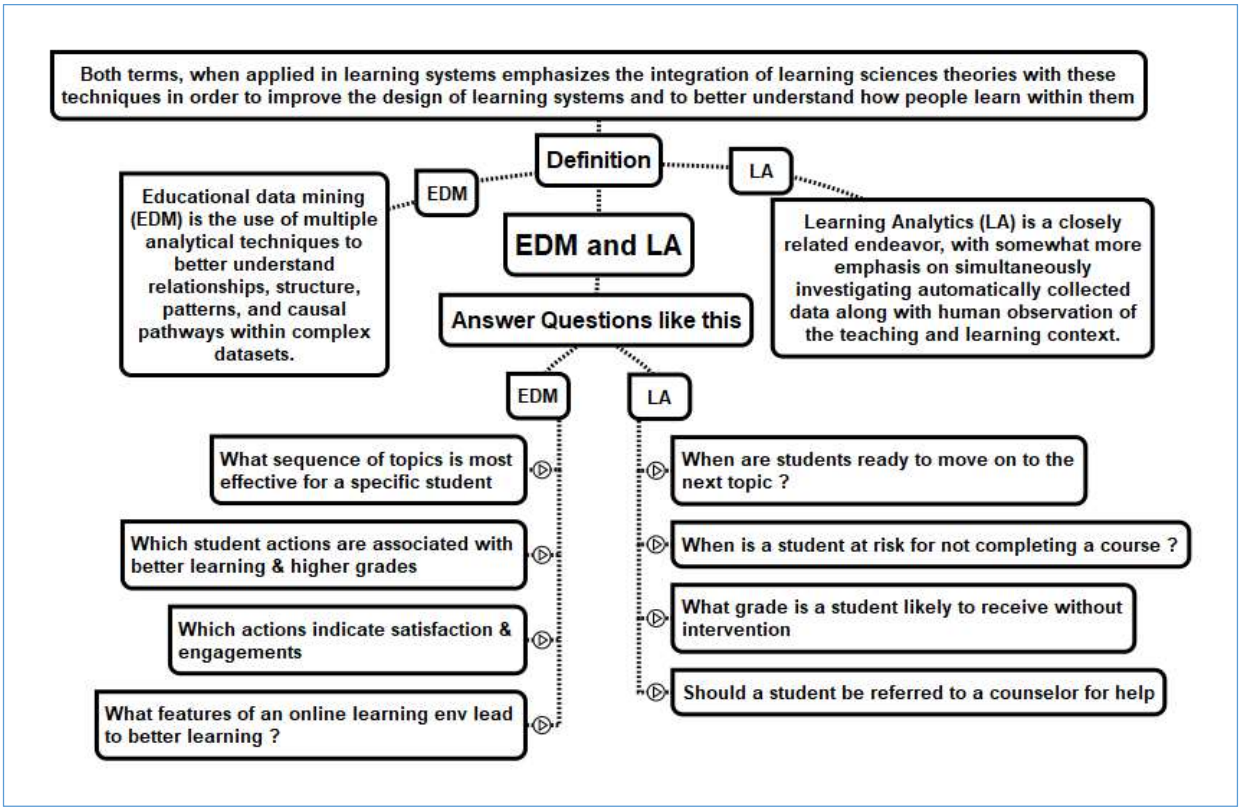

Figure 1 : Comparing EDM and LA

Data is stored in databases and from this evolved a term Knowledge Discovery in Databases (KDD) which refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods (Dutt, Ismail, & Herawan, 2017). It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The goal of the KDD process is to extract knowledge from data in the context of large databases. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing,

sampling, and projections of the data prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process. The Cross Industry Standard Process for Data Mining (CRISP-DM) is a cyclic process for development and analysis of data mining models (Leventhal, 2010). With increased demand of data mining, more algorithms have been built. This standard spell out practices which everyone can follow. It has six phases. KDD process and CRISP-DM is shown in figure 2.

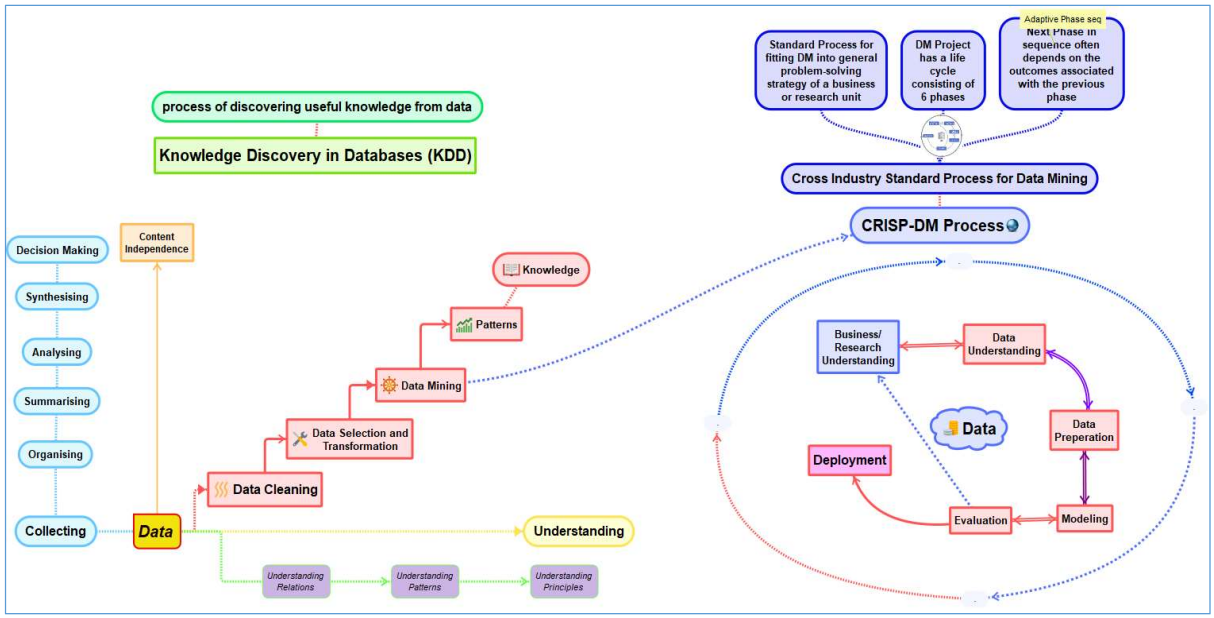

Figure 2 : KDD and CRISP model

DATA MINING TOOLS AND TECHNIQUES

Educational Data Mining (EDM) is one of the emerging fields in the pedagogy and andragogy paradigm. It concerns the techniques which research data coming from the educational domain. It includes the transformation of existing, and the innovation of new approaches derived from multidisciplinary spheres of influence such as statistics, machine learning, psychometrics, scientific computing etc. (Kamath & Kamat, 2016). This book is based on the research work undertaken by the authors on the theme "Mining of Educational Data for the Analysis and Prediction of Students' Academic Performance" using R programming.

In the paper (Roy & Garg, 2017), authors have described open source tools used for EDM and LA as – R, WEKA, Orange, Tanagra. Python, SAS and SPSS are other tools but all of them are open source. The selection of tool depends upon being – opensource/ licensed, interoperability, features with extended functionality with additional libraries, easy interface, handle large data and programming environment which can be managed easily. Integrating GitHub or other repository managements system make project management simpler to handle.

Common data mining techniques are – prediction, classification, decision tree, clustering, association rule analysis.

- Prediction. Predicting the performance of a student is a great concern to the managements of higher educational (Ramesh, Parkavi, & Ramar, 2013). Data collected at the time of admission can be used for classifying and predicting students' behavior and performance as well as teachers performance (Joseph & Devadas, 2015)

- Classification. Classification is used to predict values from other variables (Cheng, 2017). Classification maps the data into predefined sets or groups of classes. This is reason classification is categorized as supervised learning. This technique often employs, decision trees, random forests and neural network algorithms. (Roy & Garg, 2018) . This technique can identify students at risk and thereby decreasing student dropout. (Thilagaraj & Sengottaiyan, 2017)
- Clustering. This is one of the pre-processing algorithm and follows an unsupervised approach for analysing data. It groups the data into clusters which have homogeneity (Dutt et al., 2017) .
- Association Rule Mining technique determines interesting relations between attributes in data. It mainly intended to recognize strong rules using different support and confidence. (Borkar & Rajeswari, 2014). Clustering helps in grouping datasets based on similarities. It can also be used group similar course materials or grouping students based on their learning behavior patterns. (Cheng, 2017).

Table 1 : Summarizing EDM techniques for particular sample data structure

Data Mining Technique	EDM Tasks	Algorithms	R Functions / Libraries
Descriptive	Mean Marks, Category wise, Distribution of marks		mean, cor, histogram, density, ggplot2,
Prediction	Predict final marks on basis of attendance, assignment, unit test and gender.	Linear regression	lm, predict, plot,
Classification	Predict if student will pass or fail in the subject	Logistic regression, Decision Tree (CART, ctree)	glm, predict, rpart, rpartViz
Clustering	Group students on the basis of numerical values	Kmeans	kmeans
Association Rule	Finding association : Selection of electives by students	Arules	Arules, arulesViz

EDM USING R PROGRAMMING

R programming when used with R Studio IDE make data analysis easier to adopt. R comes with base packages and choice of adding functionality from more than 12,000 packages from its site (https://cran.r-project.org) . Table 1 summarizes data mining technique, EDM tasks, algorithms and R functions/ libraries which can be used for analysis on Final Grades on Semesters. Table 2 summarizes the steps of doing prediction, classification and association rule analysis for sample data and how they can be used for various educational applications.

DISCUSSION

Choosing what data to mine and how to analyses may be challenge. Student while accessing digital system may be at risk for their personal data. This may discourage them to participate digitally (Cheng, 2017) . It is important to consider protecting individual privacy. Since this is developing area, it is still not clear which visualizations techniques will best information to stakeholders for decision making. Authors in the paper (Castro, Garcia, Prata, Lisboa, & Prata, 2017) while doing exploratory study showed that different algorithms and methods are used in data mining for education. They found that decision tree was the most common technique because of the visualization it provides for better understanding.

Table-2: Steps of doing Data Mining Techniques in R and their applications for Education

Prediction (Linear Regression)	**Classification (Decision Tree)**	**Association Rule**
Data : DV – btech (numeric) ; IV – gender, cat,attnd, class10, class12	Data : DV – finalresult(binary class) ; IV – age, gender, subject1 & subject2 marks	Data : Transaction Format : choice of electives by students
Predict btech marks	Classify student pass or fail	Find Rules of choices

#data for Linear Modeling
str(data1)
head(data1)
 	
names(data1)
#Linear Modeling
model1 = lm(btech ~ . , data=data1)
summary(model1)
#keep only significant variables
model2 = lm(btech ~ attnd + class12, data=data1)
summary(model2)
 	
#verifying Model Assumptions
plot(model2)
 	
#Predict
ndata1= data.frame(attnd = c(.70,.80),class12 = c(.60, .75))
p1 = predict(model2, newdata= ndata1, type='response')
cbind(ndata1, p1)
	#libraries
	Library(rpart);
	library(rpart.plot) ;
	library(dplyr)
	
	#data for Decision Tree
	str(data1)
	head(data1)
	#create model
	dtree1 = rpart(finalresult ~ . , data= data1)
	dtree1
	printcp(dtree1) #complexity parameter
	#plot the tree
	rpart.plot(dtree1, extra=104, nn=T, main='Decision Tree to Predict Result Class- Pass/ Fail')
	#Predict for sample data
	ndata1 = dplyr::sample_n(data1, 2)
	p1=predict(dtree1, newdata=ndata1, type='class')
	cbind(ndata1, p1)
		#libraries
		library(arules); library(arulesViz)
		#data in transaction format
		summary(data1)
		inspect(data1)
		
		itemFrequencyPlot(data1, type='absolute')
		#makerules
		rules1 <- apriori(data1, parameter = list(maxlen=3, support=0.04, confidence=0.6, ext=TRUE))
		inspect(rules1)
		
		#sort rules
		rules1L = sort(rules1, by='lift', decreasing=T)
		inspect(rules1L)
		
		#specific rules :
		rules1S1 <- subset(rules1, subset=(rhs %in% "elective6") & (lift>1.0))
		inspect(rules1S1)
		#find rules with high lift, high confidence and support – **Interesting rules**

Predict Marks, Performance	Predicting binary classification – Pass/ Fail; selection in placement process	Find association between selection of electives; association between attendance and assignment etc; association between skills for placement

CONCLUSION

EDM is a field when used for solving educationally related problems. EDM provides many challenges. It is incremental in nature, variety and forms of data make interoperability difficult. Due to its nature of activity it employs other disciplinary areas like psychology, teaching and learning methodologies. Various techniques have evolved with DM algorithms but is has been found that prediction and classification are mostly used. R programming with its statistical base provides an excellent tool to implement EDM on the data generated in educational institutions.

REFERENCES

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), 3–16. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *Washington, DC: SRI International*, 1–57. https://doi.org/10.2991/icaiees-13.2013.22.

Borkar, S., & Rajeswari, K. (2014). Attributes selection for predicting students' academic performance

using education data mining and artificial neural network. *International Journal of Computer Applications*, *86*(10), 25–29.

Castro, A., Garcia, L., Prata, D., Lisboa, M., & Prata, M. (2017). An Exploratory Study on Data Mining in Education: Practiced Algorithms and Methods. *International Journal of Information and Education Technology*, *7*(5), 319–323. https://doi.org/10.18178/ijiet.2017.7.5.888.

Charitopoulos, A., & Rangoussi, M. (2017). Educational data mining and data analysis for optimal learning content management Applied in moodle for undergraduate engineering studies, (April), 990–998. https://doi.org/10.1109/EDUCON.2017.7942969.

Cheng, J. (2017). Data-Mining Research in Education. Retrieved from http://arxiv.org/abs/1703.10117
Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, *5*(c), 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247.

Jindal, R., & Borah, M. D. (2013). A Survey on Educational Data Mining and Research Trends. *International Journal of Database Management Systems (IJDMS)*, *5*(3), 53–73. https://doi.org/10.5121/ijdms.2013.5304.

Joseph, S., & Devadas, L. (2015). Student's Performance Prediction Using Weighted Modified ID3 Algorithm. *International Journal of Scientific Research Engineering & Technology*, *4*(5), 2278–2882. Retrieved from www.ijsret.org.

Kamath, R. S., & Kamat, R. K. (2016). *Educational Data Mining with R and Rattle.* River Publishers. Retrieved from https://www.riverpublishers.com/book_details.php?book_id=349.

Kumar, S. (2011). Analyzing the Concepts and Techniques of Educational Data Mining for the Enhancement of Education System, *2*, 240–244.

Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. *Journal of Direct, Data and Digital Marketing Practice*, *12*(2), 137–153. https://doi.org/10.1057/dddmp.2010.35.

Lodge, J. M., & Corrin, L. (2017). What data and analytics can and do say about effective learning. *Nature Partner Journals Npj Science of Learning*, *2*(1), 5. https://doi.org/10.1038/s41539-017-0006-5.

Nithya, D. P., Umamaheswari, B., & Umadevi, A. G. (2016). A Survey on Educational Data Mining in Field of Education. Retrieved from https://www.semanticscholar.org/paper/A-Survey-on-Educational-Data-Mining-in-Field-of-Nithya- Umamaheswari/2df5ca68026fc4c93bc88ec713678c7d02e2c49c.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting Student Performance : A Statistical and Data Mining Approach. *International Journal of Computer Applications*, *63*(8), 35–39. https://doi.org/10.1504/IJTEL.2012.051816.

Roy, S., & Garg, A. (2017). Analyzing performance of students by using data mining techniques a literature survey. *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, 130–133. https://doi.org/10.1109/UPCON.2017.8251035.

Roy, S., & Garg, A. (2018). Predicting academic performance of student using classification techniques. *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, UPCON 2017*, *2018–Janua*, 568–572. https://doi.org/10.1109/UPCON.2017.8251112.

Thilagaraj, T., & Sengottaiyan, N. (2017). A Review of Educational Data Mining in Higher Education System. *Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering*, *10*, 349–358. https://doi.org/10.15439/2017R87.