IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

Survey of Technologies for Evaluation of Student Dropout Using Educational Data

Sumit Gupta[1*], Amit Ranjan[2]

[1,2] Department of Computer Science & Engineering, SRIST, Jabalpur, Rajiv Gandhi Prodyogiki Vishwavidhyalaya, Bhopal

[*]*Corresponding Author: sumitgupta262@gmail.com, Tel.: +91-93038-11911*

DOI: https://doi.org/10.26438/ijcse/v7si10.167171 | Available online at: www.ijcseonline.org

***Abstract*—** Interpretation of the dropout students and the reason behind is the most important for the universities. Due to many different reasons such as pressure, low performance, high expectations from family, faculties and individuals it is being tough to sustain for the students. Most important source of knowing the expressions of the students in these instances is their social media interactions with other students. They express their major problems on it. But this is a challenge to process such huge data and evaluating expressions from it. Data mining techniques have given a boost in such processing and application of machine learning has become boon for it. It is found that there are many such techniques available but newest techniques which are best fit in processing of expressional data is machine learning. Surveys of such techniques have become a great source of expression evaluation.

***Keywords*—** *Cloud Computing, Fault Tolerance, Virtual Machines Migration, Resource Management*

I. INTRODUCTION

Data mining is one of the most popular research areas, due to its attention among researchers in recent years. Data mining techniques have been widely applied almost in all fields to ananyse the data for classifications, predictions, decision trees, fuzzy rules and so on. There has been increasing research interest in use of data mining techniques to investigate in the ground of education.

Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods from educational environments to explore the unique type of data and those methods to better understand students in settings in which they learn [6]. Application of EDM methods has been in discovering or improving models of a domain knowledge structure and studying pedagogical support. There has been a witnessed rapid growth of EDM in order to improve the students learning process. Academic performance is crucial factor in building the students' future [7] [8]. Predicting and analysis of students' academic performance is an indispensible milestone in educational environment. In Computing education attempts have been made to predict success in programming courses [9]. The main objective of this paper is to apply the data mining algorithms such as multilayer perception, Navie Bayes, SMO, J48, REPTree to classify and predict the novice programmers in programming education.

The International working group in EDM established a yearly international conference that began in 2008 and the Journal of Educational Data Mining in 2009. An Ample of studies have been conducted on EDM in order to discover the effect of using it on students' performance [7]. A study to predict students' academic performance, the result discovered that students' university performance is depending on unit test, Assignment, attendance and graduation percentage [9]. A study on predicting students' performance in programming course revealed that the factors like students' mathematical background, programming aptitude, gender, high school mathematics grade and locality influences their programming performance [10].

SAMPLE OF DATA MINING TECHNIQUES USED FOR PREDICTION OF STUDENT PERFORMANCE

Year	Author	DM technique	Accuracy
2011	Saurabh Pal et. al	Naïve bayes	Not assigned
		Naïve bayes	82.4%
2011	Sembiring et al.	K-Means	93.7%
		DecisionTree	80.2%
2012	Edin Osmanbegovet. al	Naïve Bayes 76.48%	
		Multilayer Perception	71.2%
		J48	73.98%

2014	Vaibhav P. Vasani et. al	Naïve Bayes	86.4%
		J48	95.9%
2015	Parneet Kaur et al	Multilayer	
		Perception	75%
		Naïve Bayes	65.13%
		SMO	68.42%
		J48	69.73%
		REPTree	67.76%
2015	Amriah mohammed shahiri et al	Neural Network	98%
		Decision Tree	91%
		SVM	83%
		K-nearest Neighbour	83%
		Navie Bayes	76%
2016	Gurmeet Kaur & williamjit Singh	J48	61.53%
		Navie Bayes	63.59%

II. ABOUT EDUCATIONAL DATA MINING (EDM)

EDUCATIONAL data mining (EDM) is a field that exploits statistical, machine-learning, and data-mining (DM) algorithms over the different types of educational data. Its main objective is to analyze these types of data in order to resolve educational research issues [27]. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn [21]. On one hand, the increase in both instrumental educational software as well as state databases of student's information have created large repositories of data reflecting how students learn [143]. On the other hand, the use of Internet in education has created a new context known as e-learning or web-based education in which large amounts of information about teaching–learning interaction are endlessly generated and ubiquitously available [60]. All this information provides a gold mine of educational data [186]. EDM seeks to use these data repositories to better understand learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners. EDM has emerged as a research area in recent years for researchers all over the world from different and related research areas, which are as follows.

1) Offline education try to transmit knowledge and skills based on face-to-face contact and also study psychologically on how humans learn. Psychometrics and statistical techniques have been applied to data, like student's behavior/performance, curriculum, etc., that was gathered in classroom environments. 2) E-learning and learning management system (LMS). E-learning provides online instruction, and LMS also provides communication, collaboration, administration, and reporting tools. Web mining (WM) techniques have been applied to student's data stored by these systems in log files and databases. 3) Intelligent tutoring system (ITS) and adaptive educational hypermedia system (AEHS) are an alternative to the justput-it-on-the-web approach, trying to adapt teaching to the needs of each particular student. DM has been applied to data picked up by these systems, such as log files, user models, etc. The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice. This process does not differ much from other application areas of DM, like business, genetics, medicine, etc., because it follows the same steps as the general DM process [219]: preprocessing, DM, and postprocessing. However, it is important to note that in this paper, the term DM is used in a larger sense than the original/traditional DM definition, i.e., we are going to describe not only EDM studies that use typical DM techniques, such as classification, clustering, association-rule mining, sequential mining, text mining, etc., but also describe other approaches, such as regression, correlation, visualization, etc., which are not considered to be DM in a strict sense.Furthermore, some methodological innovations and trends in EDM, such as discovery with models and the integration of psychometric modeling frameworks, are unusual DM categories or are not necessarily seen universally as being DM.

III. EDUCATIONAL TASKS AND DM TECHNIQUES

There are many applications or tasks in educational environments that have been resolved through DM. For example, Baker suggests four key areas of application for EDM: improving student models, improving domain models, studying the pedagogical support provided by learning software, and scientific research into learning and learners; and five approaches/methods: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models. Castro et al. suggests the following EDM subjects/tasks: applications dealing with the assessment of the student's learning performance, applications that provide course adaptation and learning recommendations based on the student's learning behavior, approaches dealing with the evaluation of learning material and educational web-based courses, applications that involve feedback to both teacher and students in e-learning courses, and developments for detection of atypical students' learning behaviors.

However, as we think that there are even more possible applications, we have established our own categories for the

main educational tasks that have employed DM techniques. These categories come from different research communities (as we have previously described in Section I), and they also use different DM tasks and techniques. On one hand, we can see in Table II that the most active communities are e-learning/LMS and ITS/AEHS. On the other hand, we will see in the following sections that the most commonly applied DM tasks are regression, clustering, classification, and association-rule mining; and the most used DM techniques/methods are decision trees, neural networks, and Bayesian networks. As we can see in Fig. 2, the categories or research lines that have the most papers published are the first eight ones (from A to G with 23 or more references each), and the categories that have the fewest papers published are the last four (from H to K with less than 15 references). We think that this may be mainly due to the fact that the first eight categories are older than the last four (and so more authors have worked on these tasks), but it could also be because of the special interest in each one. For example, although social network analysis is one of the newest tasks, it has more papers than the other three.

We also want to point out that we have organized these categories by grouping them near the most closely related ones, which in our opinion are the following: since tasks A and B provide information to instructors and C to the students; D, E, F, and G tasks reveal students' characteristics; H and I study graphs and relationships between students and concepts, respectively; and J and K help in creating/planning courseware and the course, respectively. Next, we are going to describe in detail these tasks/categories and the most relevant studies. But, as there are closely related areas, some references could be located in a different category or in several. A. Analysis and Visualization of Data The objective of the analysis and visualization of data is 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. Used with educational data, this descriptive analysis can provide such global data characteristics as summaries and reports about learner's behavior.

It is not surprising that teachers prefer pedagogically oriented statistics (overall success rate, mastery levels, typical misconceptions, percentage of exercises tackled, and material read) that are easy to interpret. On the other hand, teachers find the fine-grained statistics in log data too cumbersome to inspect or too time-consuming to interpret. Statistical analysis of educational data (logs files/databases) can tell us things such as: where students enter and exit, the most popular pages, the browsers students tend to use, and patterns of use over time,; the number of visits, origin of visitors, number of hits, and patterns of use throughout various time periods; number of visits and duration per quarter, top search terms, and number of downloads of e-learning resources; number of different pages browsed and total time for browsing different pages; usage summaries and reports on weekly and monthly user trends and activities; session statistics and session patterns; statistical indicators on the learner's interactions in forums [5]; the amount of material students might go through and the order in which students study topics; resources used by students and resources valued by students; the overall averages of contributions to discussion forums, the amount of posting versus replies, and the amount of learner-to learner interaction versus learner-to-teacher interaction; the time a student dedicates to the course or a particular part of it; the learners' behavior and time distribution and the distribution of network traffic over time; and the frequency of studying events, 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 the student has worked, how many minutes he has worked today, how many problems he has resolved, and his correct percentage, our prediction of his score, and his performance level. Information visualization uses graphic techniques to help people to understand and analyze data. Visual representations and interaction techniques take advantage of the human eye's broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once. There are several studies oriented toward visualizing different educational data such as: patterns of annual, seasonal, daily, and hourly user behavior on online forums; the complete educational (assessment) process; mean values of attributes analyzed in data to measure mathematical skills; tutor–student interaction data from an automated reading tutor; statistical graphs about assignments complement, questions admitted, exam score, etc.; student tracking data regarding social, cognitive, and behavioral aspects of students; student's attendance, access to resources, overview of discussions, and results on assignments and quizzes; weekly information regarding students' and groups' activity; student's progression per question as a transition between the types of questions; fingertip actions in collaborative learning activities; deficiencies in a student's basic understanding of individual concepts and higher education student-evaluation data; student's interactions with online learning environments; the students' online exercise work, including students' interactions and answers, mistakes, teachers' comments, etc.; questions and suggestions in an adaptive tutorial; navigational behavior and the performance of the

learner; educational trails of Web pages visited and activities done; and the sequence of LOs and educational trails. B. Providing Feedback for Supporting Instructors The objective is to provide feedback to support course authors /teachers /administrators in decision making (about how to improve students' learning, organize instructional resources more efficiently, etc.) and enable them to take appropriate proactive and/or remedial action. It is important to point out that this task is different than data analyzing and visualizing tasks, which only provide basic information directly from data (reports, statistics, etc.). Moreover, providing feedback divulges completely new, hidden, and interesting information found in data. Several DM techniques have been used in this task, although association-rule mining has been the most common. Association-rule mining reveals interesting relationships among variables in large databases and presents them in the form of strong rules, according to the different degrees of interest they might present.

IV. EXISTING SYSTEM

In past years the number of student dropout from the educational institute is increasing rapidly. The high rate of student's dropout in a registered course has been a major threat to many educational institutions or universities. The student enters the institution with lots of dreams and expectations. When their expectations are not fulfilling or certain factors like demographics will effect and makes them drop from their registered program. It is a major threat to all educational institution. The various technique of the dimensionality reduction, which includes feature selection and feature extraction. Feature selection is step by step procedure that is used to select the right attribute from a given attribute sets. For the feature extraction process, it involves the transformation of higher dimensions' data in corresponding lower dimensions. Feature selection consists of factors like Academics, demographical factors, psychological factors, health issues, teacher's opinion, student behavior. In this paper, we introduce a methodology to predict the student dropout using Naive-Bayes Classification Algorithm in R language. And also examine the reason for student drop out at an early state and predict whether the student will drop or not. There are many factors that affect a student to commit dropout as we mentioned above. Early dropout prediction helps the organization to retain the students from the respective academic program. [1] Educational Data Mining (EDM) is a research area with focus on the use of data mining algorithms/techniques in educational data, with a diversified range of purposes. Among them, EDM can be useful for inducing a model to forecast students' final performance, early in the term, in time to trigger the use of educational recovery techniques, in an attempt to prevent students' failures. This paper presents and discusses the results of three experiments on forecasting students' performance, based on real data, extracted from stored students' performance records related to an university course. [2]

With the rapid development of the Internet and communication technology, online education has drawn more and more attention, online learning platforms, on the other hand, store massive learner behavioral data and educational data. How to effectively analyze and utilize the data to improve the quality of online education has become a key issue urgently needed to be solved in the field of big data in education(BDE), educational data mining(EDM) is exactly an effective and practical method and means of applying BDE. Therefore, EDM is an important academic research hotspot in the field of EDM. Firstly, the paper introduces the basic concepts of BDE, EDM and online learning platform, and then elaborates on the process of how educational data mining transforms raw data into knowledge. Finally, the key technologies of data mining are classified according to their uses, and give its application in online education scene. The paper can provide some guidance for the research and application of educational data mining based on online education. [3]

As a student embarks on an educational journey, several factors influence his/her behavior and overall class performance. This paper presents the capabilities of Educational Data Mining (EDM), specifically the use of Association Rule Mining and Pattern Discovery, in the context of a Higher Educational Institute. Real student data comprising of parameters such as class performance, note taking, attention and self-making of assignments have been collected and examined to uncover associations that have been presented in this paper. It was found that student class performance is directly influenced by the attention given to a lecture, proper note-taking and the tendency to self-solve assignments. [4]

Data mining is the way of extracting the useful information, patterns from large volume of information by using various techniques. It is a powerful technology with great potential to help businesses to make full use of the available data for competitive advantages. This paper discusses various machine learning techniques and the detailed processes of Knowledge Discovery in Databases (KDD). This study also focus on various DM/ML approaches such as Classification, Clustering and Regression and discuss different types of each approach with its advantages and disadvantages. [5]

V. CONCLUSION

This paper is a review of the state of the art with respect to EDM and surveys the most relevant work in this area to date. In fact, after first collecting and consulting all the published bibliography in EDM area, we have selected each author's most important studies. Then, we have classified each study

not only by the type of data and DM techniques used, but also and more importantly, by the type of educational task that they resolve. EDM has been introduced as an upcoming research area related to several well-established areas of research, including e-learning, AH, ITSs, WM, DM, etc. We have seen how fast EDM is growing as reflected in the increasing number of contributions published every year in international conferences and journals and the number of specific tools specially developed for applying DM algorithms in educational data/environments. Therefore, it could be said that EDM is now approaching its adolescence, i.e., it is no longer in its early days, but is not yet a mature area. In fact, we have described some interesting future lines, but for it to become a more mature area, it is also necessary for researchers to develop more unified and collaborative studies instead of the current plethora of multiple individual proposals and lines. Thus, the full integration of DM in the educational environment will become a reality, and fully operative implementations (both commercial and free) could be made available not only for researchers and developers, but also for external users.

REFERENCES

[1] V. Hegde and P. P. Prageeth, "Higher education student dropout prediction and analysis through educational data mining," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 694-699. doi: 10.1109/ICISC.2018.8398887

[2] M. C. Nicoletti, M. Marques and M. P. Guimaraes, "A data mining approach for forecasting students' performance," 2018 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, 2018, pp. 1-7. doi: 10.23919/CISTI.2018.8399389

[3] W. Zhang and S. Qin, "A brief analysis of the key technologies and applications of educational data mining on online learning platform," 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), Shanghai, 2018, pp. 83-86. doi: 10.1109/ICBDA.2018.8367655

[4] S. R. Guruvayur and R. Suchithra, "A detailed study on machine learning techniques for data mining," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 1187-1192. doi: 10.1109/ICOEI.2017.8300900

[5] A. F. Meghji, N. Ahmed Mahoto, M. A. Unar and M. Akram Shaikh, "Analysis of Student Performance using EDM Methods," 2018 5th International Multi-Topic ICT Conference (IMTIC), Jamshoro, 2018, pp. 1-7. doi: 10.1109/IMTIC.2018.8467226

[6] C. Romero, S. Ventura, "Educational data mining : a review of the state of the art", IEEE Transactions on Systems, Man, and Cybernetics, (Applications and reviews), Vol. 40, Issue 6, pp 601-618, 2010.

[7] R.S. Baker, A.T. Corbett,K.R. Koedinger, "Detecting Student Misuse of Intelligent Tutoring Systems". Proceedings of the 7th International Conference on Intelligent Tutoring Systems, pp 531-540, 2004.

[8] T. Tang, G. McCalla," Smart recommendation for an evolving elearning system: architecture and experiment", International Journal on E-Learning, vol. 4, issue 1, pp 105-129, 2005.

[9] M. de Raadt, M. Hamilton, R.F. Lister, J. Tutty, B. Baker,I. Box, & M. Petre. "Approaches to learning in computer programming students and their effect on success". Research and Development in Higher Education Series, Vol. 28, pp 407-414, 2005.

[10] Saurabh Pal, "Data Mining: A Prediction For Performance Improvement Using Classification", International Journal of Computer Science and Information Security, Vol. 9, Isuue 4, pp 136- 140, 2011. Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9

[11] Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601–618. doi:10.1109/tsmcc.2010.2053532

Authors Profile

Mr. Sumit Gupta pursed Bachelor of Engineering in Information Technology from Rajiv Gandhi Prodyogiki Vishwavidhyalaya Bhopal in 2010, and currently pursuing Master of Technology in Computer Science from Shri Ram Institute of Science and Technology Jabalpur

Mr Amitranjan He is currently working as Assistant Professor in Department of Computer Science and Engineering in Shri Ram Institute of Science and Technology Jabalpur

