ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 3, Issue 1 (Jan. - Mar. 2015)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2015)

A Brief Review Status of Educational Data Mining

Sen, Umesh Kumar
Faculty, Dept. of Comp. Sc.& Application, Govt.MGM PG College
(Affiliated by Barkatullah University, Bhopal), Itarsi (M.P.) India

Abstract
Educational data mining is fully adolescent interdisciplinary area in the field of research & technology. In this area we are trying to analyze the success ratio, performance & productivity of the student by applying various techniques of the Data Mining. While applying this methodologies student will be enable to enhance their different learning process. This technique is also beneficial to educator, recommender, policy maker, instructor and stakeholder to design course curriculum according the need of students. In this paper tried to put emphasize on the different learning techniques such as offline educational system/traditional educational system, web mining/e-learning and intelligent tutorial system. By adopting all these learning techniques student & institutions could attain better enhancement and enrichment to obtain the knowledge in the field of academic curriculum. To apply the educational data mining effectively we'll use the various data mining tools & techniques such as: classification, association rule, clustering and decision tree etc.

Keywords
Educational Data Mining, Web Mining, Classification, Association Rule, Decision Tree.

I. Introduction

WHAT IS EDM?

The Educational Data Mining(EDM) community website, www.educationaldatamining.org, (Baker and Yacef (2009))[1], defines educational data mining as follows: "Educational Data Mining 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."

The EDM process convert the raw data coming from educational system 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 data mining like business, genetics, medicine, etc. because it follows the same steps as the general data mining process [221]: pre-processing, data mining and post processing. However, it is important to notice that in this paper the term data mining is used in a larger sense than the original/traditional DM definition. There is pressure in higher educational institutions to provide up to date institutional effectiveness (C. Romero & Ventura, 2010). The recent literature related to educational data mining (EDM)data mining which is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data (Baker & Yacef, 2009). Researchers within EDM focus on topics ranging from using data mining to improve institutional effectiveness to applying data mining in improving student learning process. EDM has emerged as a research area in recent years for researchers all over the world from different and related research areas such as:

- Offline education system/traditional educational system 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 behavior/performance, curriculum, etc. that was gathered in classroom environment.
- Web mining/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 data stored by these systems in log files and databases.
-Intelligent Tutoring system(ITS) and Adaptive Educational Hypermedia System (AEHS) are an alternative to the just-put-it-on-the-web approach by trying to adapt teaching to the needs of each particular student. Data Mining has been applied to data picked up by these systems, such as log files, user models, etc. There are various users in the field of educational data mining, who are having the different views in the context of educational research. A brief descriptions are described below in Table I.

Table.1 :Different Edm Users & Stakeholders.

Types of Users	Motive to Apply the Data Mining Techniques
Learners/ Students	To personalize e-learning; to recommend activities to learners and resources and learning tasks that could further improve their learning; to suggest interesting learning experiences to the students.
Educators/ Teachers/ Instructors	to detect which students require support; to predict student performance; to classify learners into groups; to find a learner's regular as well as irregular patterns; to find the most frequently made mistakes; to analyze students' learning and behavior; to detect which students require support.
Course Developers/Educational Researchers	to compare data mining techniques in order to be able to recommend the most useful one for each task; to develop specific data mining tools for educational purposes; etc.
System Administrators/Network Administrator	to utilize available resources more effectively; to enhance educational program offers and determine the effectiveness of the distance learning approach.

Source: The International Working Group in EDM (http://www.educationaldatamining.org) has achieved the establishment of an annual International Conference on Educational Data Mining in 2008, EDM08 [19], EDM09 [27], EDM10 [22]. This conference has evolved from previous EDM Workshops at the AIED07 [114], EC-TEL07 [224], ICALT07 [35], UM07 [17], AAII06 [34], ITS06

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

[113], AAAI05 [33], AIED05 [62], ITS04 [32] and ITS00 [30] conferences.

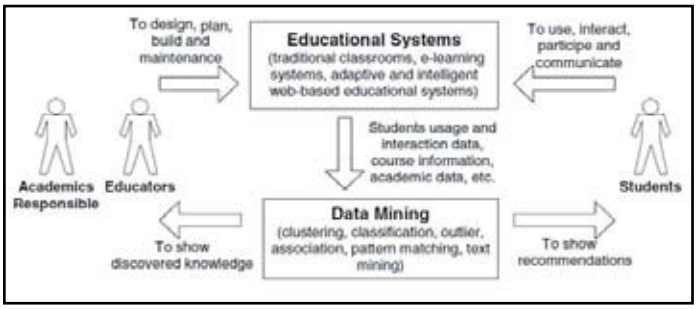

Fig: 1: The cycle of data mining in the context of educational data mining system.

II. Background & Literature Review

The educational data mining has the broad research spectrum by using the various tools & techniques of data mining . Although The number of publications about EDM has grown exponentially in the last few years .A clear sign of this tendency is the appearance of the peer-reviewed journal JEDM (Journal of Educational Data Mining) and two specific books on EDM edited by Romero & Ventura entitled: Data Mining in E-learning [222] and The Handbook of Educational Data Mining [230] co-edited with Baker & Pechenizkiy . There were also two surveys carried out previously about EDM. The first one [223] is a former review of Romero & Ventura with 81 references until 2005 in which papers were classified by the DM techniques used. The other survey [20] is a recent review by Ryan & Yacef with 46 references encompassing up to 2009. Baker and Yacef (2009) [1]defined EDM as "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".

- Z. J. Kovacic [7] presented a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success.
- Baradwaj and Pal [4] obtained the university students data like attendance, class test, seminar and assignment marks from the students' previous database, to predict the performance at the end of the semester.
- Zaïane, O. (2001). Web usage mining for a better web-based learning environment. Proceedings of Conference on Advanced Technology for Education.
- Bray [8], in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Sri Lanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socioeconomic conditions.
- Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. Proceedings of the 7th International Conference on Intelligent Tutoring Systems.
- Tang, T., McCalla, G. (2005) Smart recommendation for an evolving e-learning system: architecture and experiment, International Journal on E-Learning.

III. Knowledge Process(KDD) In Educational Data Mining

In the context of educational system, the performance and success of any student is determine by the previous semester mark, internal assessment and end semester examination. The internal assessment is carried out by the teacher based upon students performance in educational activities such as theory class, lab work, class test, presentation, seminar, assignments, communication skill, behavior and attendance etc . The end semester examination is at where score done by the student. Each student has to get minimum marks to pass a semester in internal assessment as well as end semester examination.

1. **DATA CLEANING**: Is used to remove noise and inconsistent data from sampling data. In data cleaning if users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied to it. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output.
2. **DATA SELECTION**: Where data relevant to the analysis task are retrieved from the database. The training data set was received from RGPV, Bhopal(M.P.) Based upon the sampling method by the Dept. of Computer Application(MCA) from session 2000 to 2005.Initially,the size of the taken data set is 60.All above written parameter is applied over training data set and ensured the accountability of performance of the each student.
3. **DATA TRANSFORMATION**:Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operation, for instance, attribute data may be normalized so as to fall between a small range, such as from 0.0 to 1.0. The derived variable are defined in the table. II.

Table.2: Student Related Attributes

ATTRIBUTES/ KEYWORDS	DESCRIP-TION	STANDARD NORMS
PSM	Previous Semester Mark	{First ≥ 65% Second ≥55 & <65% Third ≥ 45 & <55%, Fail < 45%}
CTG	Class Test Grade	{Poor, Average , Good}
SEMP	Seminar Performance	{Poor, Average, Good}
ASS	Assignment	{Yes, No}
ATT	Attendance	{Poor, Average, Good}
LW	Lab Work	{Yes, No}
CS	Communication Skill	{Poor, Average , Good}
ESM	End Semester Mark	{First ≥ 65% Second ≥55 & <65% Third ≥ 45 & <55%, Fail < 45%}

PSM – Previous Semester Marks/Grade obtained in MCA course. It is split into three class values: First –{First ≥ 65% Second ≥55 & <65% Third ≥ 45 & <55%, Fail < 45%}.

CTG – Class test grade obtained. Here in each semester two class tests are conducted and average of two class test are used to calculate sessional marks. CTG is split into three classes: Poor, Average , Good.

SEMP – Seminar Performance obtained. In each semester seminar are organized to check the performance of students. Seminar performance is evaluated into three classes: Poor, Average, Good.

ASS – Assignment performance. In each semester two assignments are given to students by each teacher. Assignment performance is divided into two classes: Yes – student submitted assignment, No –Student did not submit assignment.

ATT – Attendance of Student. Minimum 75% attendance is compulsory to participate in End Semester Examination. But even though in special cases low attendance students also participate in End Semester Examination on genuine reason basis. Attendance is divided into three classes: Poor - <60%, Average - ≥ 60% and < 80%, Good - ≥ 80%.

LW – Lab Work. Lab work is divided into two classes: Yes – student completed lab work, No –student did not complete lab work.

CS-While pursuing the whole session, the communication skill of the student was: Poor, Average, Good.

ESM - End Semester Marks obtained in MCA semester and it is declared as response variable. It is split into five class values: –First ≥ 65% Second ≥55 & <65% Third ≥ 45 & <55%, Fail < 45% .

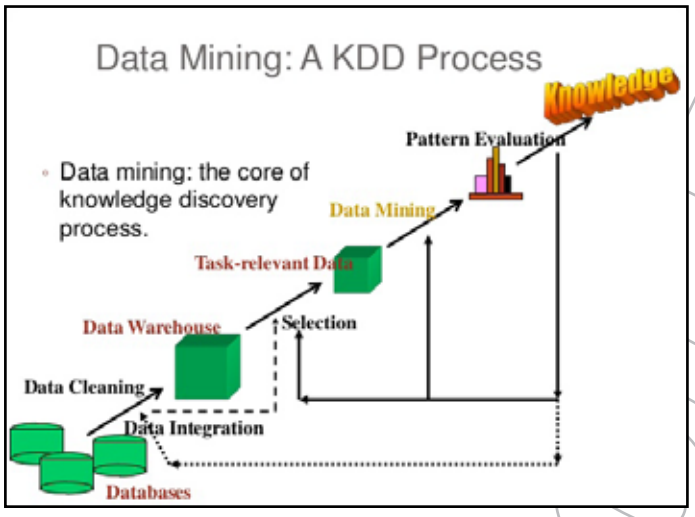

Fig.2: The Knowledge Discovery Process in Data Mining

IV. Data Miing Algorithm Over Educational System

1. **ID3 (Iterative Dichotomiser 3):** The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets. ID3 is a simple decision tree learning algorithm developed by Ross Quinlan [14].

   CART: CART [18] stands for Classification And Regression Trees introduced by Breiman. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tre Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.
2. **C4.5**: This algorithm is a successor to ID3 developed by Quinlan Ross [14]. It is also based on Hunt's algorithm.C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes.
3. **Bayes Theorem's**: The naïve Bayes approach has several advantages: It is easy to use; unlike other classification approaches only one scan of the training data is required; easily handle mining value by simply omitting that probability [11]. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters. Bayes classification has been proposed that is based on Bayes rule of conditional probability. Bayes rule is a technique to estimate the likelihood of a property given the set of data as evidence or input Bayes rule or The Bayes theorem is as follows: Let X={x1, x2, ...,xn} be a set of n attribute in Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as-

$$P(H|X) = P(X|H)\,P(H) / P$$

V. Result & Discussion

There are four classification techniques applied over training data set. They are having various results and accuracy, which mentioned in drawn table as below-

Table.3. Accuracy Through Classifier

NAME OF ALGORITHM	CORRECTLY CLASSIFIED (IN PERCENTAGE)	INCORRECTLY CLASSIFIED (IN PERCENTAGE)
ID3	55.645	32.343
CART	59.973	42.452
C4.5	48.745	51.851
NAÏVE BAYE'S	43.879	38.658

Table III. shows that a CART technique has highest accuracy of 56.25% compared to other classification algorithm.

VI. Conclusion & Future Research Work

This paper is a review of the state of the art with respect to EDM. EDM has been introduced as an up and coming research area related to several well-established areas of research including e-learning, adaptive hypermedia. Educational data mining (EDM) is an area full of exciting opportunities for researchers and practitioners. This field assists higher educational institutions with efficient ways to improve institutional effectiveness and student learning. It will be exciting to see how EDM develops over the coming years because still it is in infancy. This study would be helpful to student, teacher and institution to enhance the performance and productivity effectively. In this research paper different classification method is used to predict the performance of students.

VII. Acknowledgements

The author is gratefully acknowledge Cristobal Romero and

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Sebastian Ventura for their excellent a review of the State-of-the-Art of Educational Data Mining, which influenced my review of research paper and the field consistely.

References

[1]. *Arnold, A., Scheines, R., Beck, J.E., Jerome, B. (2005). Time and Attention: Students, Sessions, and Tasks. In AAAI2005 Workshop on Educational Data Mining, Pittsburgh.*

[2]. *Antunes, C. 2008. Acquiring background knowledge for intelligent tutoring systems. In International Conference on Educational Data Mining, Montreal, Canada.*

[3]. *Baker, R., Merceron, A., Pavilk, P.I. (2010). Educational Data Mining 1010: 3st International Conference on Educational Data Mining, Proceedings, Pittsburgh, USA.*

[4]. *Barnes, T., Desmarais , M., Romero, C., Ventura, S. (2009). Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings. Cordoba, Spain.*

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

[6]. *Dekker, G.W., Pechenizkiy, M., Vleeshouwers, J.M. (2009). Predicting Students Drop Out: A C ase Study. In International Conference on Educational Data Mining, Cordoba, Spain.*

[7]. *Dong, G., Pei, J. (2007). Sequence Data Mining. Springer.*

[8]. *Espejo, P., Ventura, S., Herrera, F. (2010) A Survey on the Application of Genetic Programming to Classification. IEEE Transactions on Systems, Man, and Cybernetics-part-C.*

[9]. *Fausett, L.V., Elwasif, W. (1994). Predicting performance from test scores using back propagation and counter propagation. In IEEE World Congress on Computational intelligence, paris, france.*

[10]. *Garcia, E., Romero, C., Ventura, S., Castro, C. (2009b). Collaborative Data Mining Tool for Education. In International Conference on Educational Data Mining, Cordoba, Spain.*

[11]. *Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.*

[12]. *Kotsiantis, S.B., Pintelas, P.E., (2005). Predicting Students' Marks in Hellenic Open University. In IEEE international Conference on Advanced Learning Technologies, Washington, DC.*

[13]. *Madhyastha, T., Hunt, E. (2009). Mining Diagnostic Assessment Data for Concept Similarity. Journal of Educational Data Mining.*

[14]. *Romero, c., ventura, s., hervás, c., gonzales, p. (2008). Data mining algorithms to classify students. In International Conference on Educational Data Mining, Montreal, Canada.*

[15]. *Romero, C. Ventura, S., Pechenizkiy, M., Baker, R. (2010). Handbook of Educational Data Mining. Taylor & Francis.*

[16]. *Shen, L., Shen, R. (2004). Learning content recommendation service based-on simple sequencing specification. In International Conference on Web-based Learning, Beijing,China.*

[17]. *Simko, M., Bielikova, M. (2009). Automatic concept relationships discovery for an adaptive e-course. In International Conference on Educational Data Mining, Cordoba, Spain.*

[18]. *Ventura, S., Romero, C., Hervas, C. (2008). Analyzing rule evaluation measures with educational datasets: a framework to help the teacher. In International Conference on Educational Data Mining, Montreal, Canada.*

[19]. *Vialardi, C., Bravo, J., Shafti, L., Ortigosa, A. (2009). Recommendation in higher education using data mining techniques. In International Conference on Educational Conference, Cordoba, Spain*

[20]. *Want, T., Mitrovic, A. (2002). Using Neural Networks to Predict Student's Performance. In International Conference on Computers in Education, Washington, DC.*

[20] *Wang, W., Weng, J., Su, J., Tseng, S. (2004). Learning portfolio analysis and mining in scorm compliant environment.In ASEE/IEEE Frontiers in Education Conference,Georgia.*

[21]. *Zafra, A., Ventura, S. (2009), Predicting student grades in learning management systems with multiple instance programming.In International Conference on Educational Data Mining, Cordoba, Spain.*

[22]. *Zaïane, O. (2002). Building A Recommender Agent fore-Learning Systems. In Proceedings of the International Conference in Education, Auckland, New Zealand.*

Author Profile

Umesh Kumar Sen, Education: Received MCA degree from RGPV, BHOPAL (M.P.), Occupation: Working as a faculty in Dept.of Computer Science & Application in Govt.MGM PG College(Affiliated by Barkatullah University,Bhopal), Itarsi (M.P.) 461111. Address: C/o Dept.of Computer Science & Application, Govt. MGM PG College, Nyas Colony, Itarsi(M.P.) 461111.