International Journal of Advances in Computer Science and its Applications – IJCSIA
Volume 4 : Issue 1 *[ISSN 2250 – 3765]*
Publication Date : 09 January 2014

Mining Educational Data : A Review on Students's Pattern of Behaviours & Performances

Wan Aezwani Wan Abu Bakar
Department of Computer Science
Faculty of Science & Technology,
University Malaysia Terengganu
21030 Kuala Terengganu,
Terengganu.
beny2194@yahoo.com,

Mohd Yazid Md. Saman
Department of Computer Science
Faculty of Science & Technology,
University Malaysia Terengganu
21030 Kuala Terengganu,
Terengganu.
yazid@umt.edu.my

Masita Abd Jalil
Department of Computer Science
Faculty of Science & Technology,
University Malaysia Terengganu
21030 Kuala Terengganu,
Terengganu.
masita@umt.edu.my

***Abstract*—The main focus of lower or higher education institutions is to provide the quality education to its students. One way to achieve the highest level of quality education system is through discovering the hidden knowledge that relied in huge educational datasets. There is an increasing trend in using data mining in education. This new emerging field, called Educational Data Mining (EDM), concerns with developing methods that discover knowledge from data originating from educational environments. Examples of educational domain are discovering knowledge for prediction regarding enrollment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the student's result sheet, prediction about student's performance and analyzing a pattern of student's class attendance and student's class absenteeism. The main objective of this paper is to study and compare the use of several data mining tasks/techniques that have been implemented mostly in the area of educational domain. By this way, the pattern of student's behavior as well as performance shall be predicted in an efficient manner.**

***Keywords*—Educational Data Mining (EDM), Knowledge Database and Discovery (KDD), Student's Behaviour, Student's Performance**

I. INTRODUCTION

There's always a need to integrate the Information and Communication Technologies (ICT) with the enrichment of teaching and learning in education. Thus many of the academic institutions are obliged to provide basic safe network for their users. In quite a short period of time, the use of ICT has vastly increased and marked a strong effect especially on school's teaching and learning.

As a result, we can have a huge amount of data flooding around companies, organizations or even individual [1]. As a matter of fact, these data itself is critical to a company's growth. It contains knowledge that could lead to the next important business decisions. Thus, this is the time where we need to have some mechanisms on keeping those data. We need to have a database.

Now, we have the storage for keeping the data and information. But, according to [2], it has been estimated that the amount of information in the world doubles every 20 months. The size and number of databases probably increases even faster. In 1989, the total number of databases in the world was estimated at five million, although most of them are small DBASE III databases. What are we supposed to do with this flood of raw data ? Clearly, little of it will ever be seen by human eyes.

Basically in many cases, these data has never been reviewed in superficial manner such that it can be seen as "*data rich but knowledge poor*" [1]. Now, it's time where when the amount of data is so enormous that human cannot process it fast enough to get the required information at the right time, then the machine learning technology has been established to potentially solve this problem. Also, there is an urgent requirement for a new generation of computational theories and tools to assist human in extracting useful information (knowledge) from the rapidly growing volumes of digital data [3]. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD).

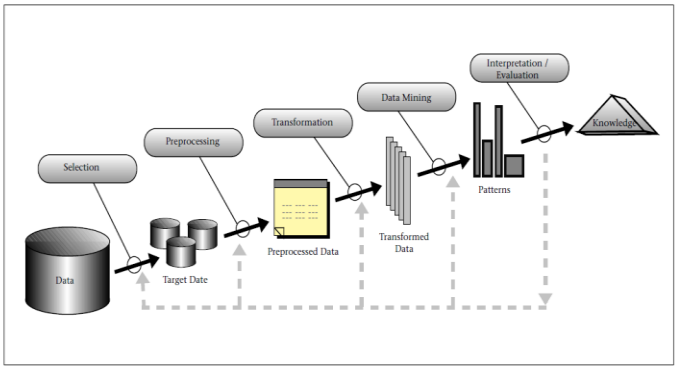

Figure 1 : Knowledge Discovery Process

Figure 1 above depicts that data mining is just a step in the KDD process that consist of applying data analysis and discovery algorithms that, under acceptable computational

efficiency limitations, produce a particular enumeration of patterns (or models) over the data. In short, the ultimate goal of knowledge discovery and data mining process is to find the patterns that are hidden among the huge sets of data and interpret them into useful knowledge and information. As a result, the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge, is in to be explored in greater depth through data mining.

The fact is that the knowledge is hidden among educational datasets and it is extractable through the use of data mining techniques. The next section outlines several related works that have been studied and compared to discover the outcome and current trends of the implementation of data mining techniques and tools which have been imposed to overcome education related problems.

II. RELATED WORKS

Data mining techniques have been applied in education and higher education. This is termed as educational data mining (EDM). Several techniques have been suggested for EDM such as association analysis, classification and prediction, clustering analysis and outlier analysis [4]. The use of **apriori algorithm** in association rule data mining has been applied to student's assessment data to improve the quality of education [5]. Tanagra [6], a free data mining software for research and education has been used as the data mining tool for the experimentation. It shows that it has a potential to use association rule mining to improve the student's performance.

Data mining techniques have been applied to higher education in India [7]. Techniques such as clustering, classification (Decision Tree) and association rules are used to improve student's performance, selection of course and major, their retention rate and for grant/fund management. The authors suggest that if data mining techniques can be applied to higher education processes, then it can help to improve students' performance from Indian perspective.

The use of educational data mining (EDM) in the field of education was also discussed in [8]. The discussion was on how data mining tools and techniques can help in enhancing the performance of educational institutions.

Decision Tree (DT) is proposed as a tool in the Student Learning Result Evaluation System (SLRES) [9]. With the model practiced, student learning can become more energetic and challenging. The author hoped to generate more theories in data mining.

Some researches combined the use of classification and clustering in data mining task in order to improve the performance of evaluation. A Student's Academic Performance Predictor [10] is proposed by using **Smooth Support Vector Machine (SSVM) classification** and kernel **K-Means clustering** technique. It is suggested that data mining techniques can be used to develop performance monitoring and evaluation tool. It was an effective improvement tool for student's performance.

Classification, one of several tasks in data mining seems to be popular and interesting among the data mining researchers. **Decision Tree** (DT) classification [11] is used as a prediction tool on a performance improvement of an engineering student where they have experimented on three DT algorithms namely **C4.5 algorithm**, **ID3 algorithm** and **CART algorithm**. The prediction outcome is the number of students who are likely to pass, fail or promoted to next year. It seems to confirm that the classification in Data mining is an interesting topic to discuss where it can accurately classifies the student's performance in First Year Engineering Course.

The classification via clustering approach is proposed to predict the final marks in a university course by taking Moodle forum system as a case study [12]. They used Weka [13], also a free data mining tool where it has showed that the **Expectation Maximisation (EM), a clustering algorithm** in Weka yielded the result similar to those of the best classification algorithm. They have suggested that the student's participation in forum was a good predictor of the final marks for the course. It types of messages and was suggested that a data text mining algorithm could be used to automatically detect and classify evaluate them.

A data mining hybrid procedure based on **Neural Network (NN)** and Data Clustering has been proposed by [14] that enabled academicians to predict student's GPA according to their foreign language performance. The authors have classified student in a well-defined cluster for further advising by using **K-Means Clustering algorithm**. The neural network tool used was Palisade [15], one of the world's leading risk and decision analysis solution software which allowed a learning pattern in a set of known data, and used those patterns to make prediction from new, incomplete data. Their future work is to study on different attribute and targeting new learning related.

A prediction on student's failure at school using **genetic programming** and different data mining approaches is proposed in [16] with high dimensional and imbalanced data. This is to obtain more comprehensible and accurate classification rule. In future, they were to carry out more experiment and data, to look at different educational level i.e. primary, secondary and higher education, to test whether the same performance results could be obtained with different DM approaches (feature selection, data balancing sand cost-sensitive classification).

Another hybrid approach which uses EDM and **regression analysis** has been proposed to analyse live video streaming (LVS) of student's online behaviour and their

performance in their course [17]. It is then suggested that there was no correlation between student's number of questions, chat messages and login times to student's success.

The demonstration on using Data Mining techniques to make prediction has been conducted on the students who are going to take the computer proficiency test and fail [18]. Three different clustering techniques are used which are **K-Means, Self-Organising Map (SOM)** and **two-step clustering (BIRCH).** After clustering result is found, the DT is used to extract useful rules from each of the identified clusters. The rules are then used to warn or counsel students who seemed to be failed. Their next project is to implement and apply the same approach to English language assessment.

An automatic inquiring system for learning materials has been presented in [19] in which, it utilized data sharing and fast searching properties of LDAP (Lightweight Directory Access Protocol). They have combined a classification and association rule in DM where the **Apriori algorithm** and **Tree-based algorithm** are employed to develop the association rule for the learning material recommendation. Association rule is adopted to find out the relation between the keywords learners used for searching the contents. And also the collaborative filtration is applied to automatically filter the correct keywords of each course. The authors planned to investigate the usability and instructional value of LMS which includes the presentation of material categories and the trouble-free search for material, the convenience of data retrieval and also the level of acceptance and comprehensive understanding of the learning materials.

The **K-Means clustering** algorithm is used to be applied in analyzing student's learning behaviour [20]. Their model has attempted to make a prediction about pass and fail ratio of students based on class performance and class attendance. The main goal of clustering is to partition students into homogeneous group according to their characteristics and abilities. The researchers hoped to refine the technique in order to get more valuable and accurate outputs.

The research in association rule has become a preferred field in data mining task. An efficient algorithm for the discovery of frequent itemset has been found [21]. Three main techniques are employed i.e. first, cluster itemset using equivalence classes or maximal hypergraph cliques. Then generate true frequent itemset from each cluster sublattice using top-down, bottom-up or hybrid lattice traversal. Experimental results have indicated that some major improvements have been attained over previous algorithms.

A new technique using classifier which is called **Fastest Association Rule Mining-Algorithm Predictor (FARM-AP)** has been presented by which it can predict the fastest ARM algorithm with 80% accuracy and very low overhead [22]. They have experimented on 3 frequent itemset mining algorithm such as **Apriori, FP-Growth** and **Eclat**. They have claimed that selecting the most appropriate algorithm for frequent itemset is also relevant to parallel computing. For future, the accuracy of FARM-AP can be increased and to train the FARM-AP with more features such as average maximal pattern length of the datasets.

A survey is performed on frequent itemset mining algorithm between **FP-Growth, Eclat**, **RElim (Recursive Elimination) algorithm** and **SaM (Split and Merge) algorithm** [23]. They have tested on 4 different datasets which are obtained from UCI (University of California at Irvine) repository. They have mentioned that different algorithm outperforms differently on different datasets. But it can be considered that SaM is an improved performer from the overall dataset analysis.

The different DM tasks are compared in an attempt to find the best performer. A comparison between traditional classification algorithms versus classification based on association rule (AR) algorithm is then presented with regards to classification accuracy, number of derived rules, rules features and processing time [24]. They have stated that the classification and association rule discovery are similar except that classification involved prediction of one attribute (i.e the class) while association rule discovery can predict any attribute in the datasets. The **traditional classification algorithms** that they have tested **are OneR, DT C4.5, PART, Naive Bayes** and **RIPPER**. They compared with the two well known **classification based on association rule algorithms** namely **CBA** and **MMAC (Multi-class, Multi-label Associative Classification)** [25]. They have discovered that MMAC performed consistently well in term of classification accuracy on both artificial dataset and real world dataset. For future work, they are looking on text mining classification and how to extract negative association rule which basically represents inferring items from the absence of other items in the customer shopping cart.

Data mining task has been applied in manufacturing industry [26]. The DM prediction algorithms (i.e. **CHAID, C&R, Quest, Neural Network, Bayesian, logistic regression and SVM)** have been applied on detecting a wasted parts of car aluminium parts (i.e. engine bracket). The engine bracket is belonged to Ahanpishegan Company, a manufacturer in automotive industry. They aimed to improve industrial product reliability, maintainability and availability. They have recommended on using data stream mining to achieve quicker and more accurate results in the future.

The AR mining algorithm is implemented on **Apriori** and **Eclat,** where both algorithms are known for the best basic algorithms in mining frequent itemset in a set of transactions [27]. These algorithms use several optimizations to achieve maximum performance with regards to both execution time and memory usage. The Apriori implementation is based on prefix tree representation of the needed counters and uses a doubly recursive scheme to count the transactions. The Eclat implementation uses bit matrices to represent transactions lists

and to filter closed and maximal itemsets. It is stated that the main difference between Apriori and Eclat is how they traverse this prefix tree and how they determine the support of an itemset (i.e. the number of transactions the itemset is contained in). Apriori traverses the prefix tree in breadth first order that is it first checks itemset of size 1, size 2 and so on. Eclat traverses in depth first order that is it extends an itemset prefix until it reaches the boundary between frequent and infrequent itemsets and then backtracks to work on the next prefix (in lexicographic order w.r.t. the fixed order of the items). The author has experimented on five different datasets (i.e. BMS-Webview-1, T10I4D100K, census, chess and mushroom). The result has shown that Eclat won the competition to four out of five datasets w.r.t. execution time and always won w.r.t. memory usage. If the number of maximal itemset is high, Apriori won due to its efficient filtering while Eclat won for a lower number of maximal itemset due to its more efficient search.

Refer to **Table 1** on **Appendix A** for the frequency checklist and summary of the DM algorithms used in the literature reviews.

III. CONCLUSION AND FUTURE WORKS

As stipulated from the literatures, it is quite clear that there are no established criteria for deciding which methods to use in which circumstance, and many of the approaches are based on crude heuristic approximations to avoid the expensive search required to find optimal, or even good solutions [3].

Mining the association rules (ARs) can be classified as one of the most popular and prominent areas in data mining. It aims at discovering interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [28]. Interestingly, it continues to become an active research in knowledge database and discovery (KDD) [29]. In our next paper, we will focus on the three AR techniques namely Apriori, FP-Growth and Eclat. We will try to enhance these algorithms with our proposed algorithm and to test into several available datasets.

ACKNOWLEDGMENT

We wish to thank Mustafa Man for his insightful comments and suggestions.

REFERENCES

[1] Data mining, available at http://www.zentut.com/data-mining/data-mining-processes (Retrieved on 28/10/2012)

[2] W.-J. Frawley, G. Piatetsky-Shapiro, & C.-J. Matheus, *Knowledge Discovery in Databases : An Overview*, AI Magazine, Vol. 13, No. 3, (1992).

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases*, AI Magazine Vol. 17 No. 3, (1996).

[4] V. Kumar, and A. Chadha, *An Empirical Study of the Applications of Data Mining techniques in Higher Education*, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 3, (2011).

[5] V. Kumar, and A. Chadha, *Mining Association Rules in Student's Assessment Data*, International Journal of Computer Science Issues (IJCSI), Vol. 9, Issue 5, No. 3, (2012).

[6] Tanagra, available at http://eric.univ-lyon2.fr/~ricco/tanagra/ (Retrieved on 28/12/2013)

[7] J. Ranjan, and R. Ranjan, *Application of Data Mining Techniques In Higher Education In India*, Journal of Knowledge Management Practice, Vol. 11, Special Issue 1, (2010).

[8] P. Gulati, & A. Sharma. *Educational Data Mining for Improving Educational Quality*, International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN : 2249 – 9555, Vol. 2, No. 3, (2012).

[9] H. Sun. *Research on Student Learning Result System Based on Data Mining*, International Journal of Computer Science and Network Security, Vol. 10, No. 4, (2010).

[10] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, & E. Wani. *Prediction of Student Academic Performance By An Application of Data Mining Techniques*, International Conference on management and Artificial Intelligence (IPEDR), Vol. 6, IACSIT Press, Bali, Indonesia, (2011).

[11] Surjeet K. Yadav, & S. Pal. *Data Mining : A Prediction for Performance Improvement of Engineering Student Using Classification*, World of Science and Information Technology Journal, ISSN : 2221 – 0741, Vol. 2, No. 2, pp. 51 – 56, (2012).

[12] M. I. Lopez, J. M. Luna, C. Romero, & S. Ventura. *Classification Via Clustering for Predicting Final Marks Based On Student Participation In Forums*, (2012).

[13] Weka Download, available at http://www.cs.waikato.ac.nz/ml/weka/downloading.html (Retrieved on 28/11/2013)

[14] C. E. Moucary, M. Khair, & W. Zakhem. *Improving Student's Performance Using Data Clustering and Neural Networks In Foreign-Language Based Higher Education*, The Research Bulletin of Jordan ACM, Vol. 11 (III), p. 27–34, (2012).

[15] Neural Network Tool, available at http://blog.palisade.com/blog/neural-network (Retrieved on 21/2/2013)

[16] C. Marquez-Vera, A. Cano, C. Romero, & S. Ventura. *Predicting Student Failure At School Using Genetic Programming and Different Data Mining Approaches With High Dimensional and Imbalanced Data*, © Springer Science + Business Media, LLC, (2012).

[17] M. Abdous, W. He, & C-J. Yen. *Using Data Mining for Predicting Relationships Between Online Question Theme and Final Grade*. Educational Technology & Society, 15 (3), pp. 77 – 88, (2012).

[18] C.-F. Tsai, C.-T. Tsai, C.-S. Hung, & P.-S. Hwang. *Data Mining Techniques For Identifying Students At Risk of Failing Computer Proficiency Test Required For Graduation*, Australasian Journal of Educational Technology, 27(3), pp. 481 – 498. (2011).

[19] L. Feng-Jung & S. Bai-Jiun. *Application of Data Mining Technology on E-Learning Material Recommendation*, E-Learning Experiences & Future, Safeeullah Soomro (Ed.), ISBN : 978-953-307-092-6, InTech. (2010).

[20] S. Ayesha, T. Mustafa, A.-R. Sattar, & M.-Inayat Khan, *Data Mining Model for Higher Education System*, European Journal of Scientific Research, ISSN : 1450-2160X, Vol. 43, No. 1, pp. 24-29, (2010).

[21] M.-J. Zaki, S. Parthasarathy, M. Ogihara, & W. Li. *New Algorithm for Fast zdiscovery of Association Rule*, 3rd International Conference on Knowledge Discovery and Data Mining, (1997).

[22] M. HooshSadat, H.-W. Samuel, S. Patel, & O.-R. Zaiane. *Fastest Association Rule Mining Algorithm Predictor (FARM-AP),* ACM 978-1-4503-0626-3/11/05, (2011)

[23] S. Pramod, & O.-P. Vyas. *Survey on Frequent Itemset Mining Algorithms*, International Journal of Computer Applications (0975-8887), Vol. 1, No. 5, (2010).

[24] A. Al Deen, M. Nofal, & S. Bani-Ahmad. *Classification Based On Association Rule Mining Techniques : A General Survey and Empirical Comparative Evaluation*, Ubiquitous Computing and Communication Journal, Vol. 5, No. 3, pp. 9-17, (2011).

[25] F. Thabtah, P. Cowling, & Y. H. Peng, *MMAC : A New Multi-Class, Multi-Label Associative Classification Approach*, 4th IEEE International Conference on Data Mining (ICDM 04). (2004).

[26] G.Amooee, B. Minaei-Bidgoli, & M. Bagheri-Dehnavi, *A Comparison Between Data Mining Prediction Algorithms For Fault Detection (Case Study : Ahanpishegan Co.*, International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 6, No. 3, (2011).

[27] C. Borgelt, *Efficient implementations of Apriori and Eclat*. Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI), Melbourne, Florida. (2003).

[28] R. Agrawal, T. Imielinski & A. Swami. *Database Mining : A Performance Perspective*, IEEE Transactions on Knowledge and Data Engineering, 5(6), p. 914 (1993).

[29] Z. Abdullah, T. Herawan & M.M. Deris. *Detecting Critical Least Association Rules in Medical Databases*, International Journal of Modern Physics : Conference Series, Vol. 1, No. 1, pp. 1-5, (2010).

APPENDIX A

Table 1 : The frequency checklist and summary of the DM algorithms used in the literature reviews

Author (Reference No.)	Classification															ARM					Clustering				Classification based on ARM	
	SVM	SSVM	C4.5	ID3	CART	NN	GP	RA	OneR	PART	NB	RIPPER	CHAID	C&R	Quest	Apriori	Eclat	FP-Growth	SaM	RElim	K-Means	EM	SOM	BIRCH	CBA	MMAC
[18]																√										
[22]		√																			√					
[23]			√	√	√																					
[24]																							√			
[25]						√															√					
[26]							√																			
[27]								√																		
[28]																					√		√	√		
[29]																√										
[30]																					√					
[31]																	√									
[32]																√	√	√								
[33]																	√	√	√	√						
[34]			√						√	√	√	√													√	√
[36]	√					√		√			√		√	√	√											
[37]																√	√									

Note :
SVM – Support Vector Machine
SSVM – Smooth Support Vector Machine
NN – Neural Network
GP – Genetic Programming
RA _ Regression Analysis
NB – Naive Bayes
FP Growth – Frequent Pattern Growth
SaM – Split and Merge
RELim – Recursive Elimination
EM – Expectation Maximisation
SOM – Self Organising Map