Research in Education with Data- Mining

Khin Htay[1,*], Mie Mie Aung[2], Yin Cho[2], Moe Moe Thein[3]

[1,*]Faculty of Information Science

[2]Faculty of Information Technology Supporting and Maintenance

[3]Faculty of Computer System and Technology

[1*,2,3]University of Computer Studies (Meiktila), Myanmar

khinhtay.ucs@gmail.com

Abstract: *Data mining (DM) is popular in education area especially when examining student's learning performances. DM does help education institutions provide high- quality education for its leaners. Applying data mining in education also known as educational data mining (EDM), which enables to better understand how student learn and identify how improve educational outcomes. This paper is designed to justify the capabilities of data mining approaches in the field of education. There exit various methods and application in EDM which can follow both applied research objectives such as improving and enhancing learning quality, as well as objective tend to improving our understanding of the learning process. Several specific algorithms, methods, applications and gaps in the current literature and future insights are discussed here.*

 Keywords: Educational Data Mining (EDM); Data Mining (DM); Algorithm; Clustering; Classification; Regression.

Introduction

 There is pressure in higher educational institutions to provide up-to-date information institutional effectiveness. Education institutions would like to know, for instance, which students will enroll in particular course programs, and which students will need assistance fir graduation. Though the analysis and the presentation of data they collected, or data mining, they changes of these students or learners are able be effectively addressed. Data mining enables organizations to uncover and understand hidden patterns in vast databases by using their current reporting capabilities. And these patterns are then built into data mining models and applied to predict individual behavior and performance with high accuracy. In this way, resources and staff can be allocated by institutions with an accurate estimate of how may students will take action before he or she drops out.

Educational data mining (EDM) is an emerging discipline including but not limited to information retrieval, recommender systems, visual data analytics, social network analysis(SNA), cognitive psychology, psychometrics, and so on. Its methods is often different from those methods from the broader data mining literature. EDM draws from several reference disciplines including data mining., learning theory, data visualization, machine learning psychometrics[2] And this emerging field of EDM examines the unit ways of applying data mining techniques to solve educationally related issues.

This paper is to synthesize and share various example by using data mining in education, to support reflection on teaching and learning. The background of EDM is described, then various algorithms that frequently used are briefly presented. Some application demonstrate how data mining are used to save resources and help teachers and learners. Finally, we conclude the paper.

BACKGROUND OF DATA MINING

Data mining is an interdisciplinary subfield of computer science[3-5]. Data mining is The analysis step if the "knowledge discovery in database" process, or KDD [6].

Data mining techniques have their roots in machine learning, artificial intelligence, computer science, and statistic etc.[7]. And data mining is an exploratory process, but it can be use for confirmatory investigation [8]. It is different from other searching and analysis techniques because data mining is high exploratory, where other analyses are typically problem-driven and confirmatory. Through the combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge, hidden trends and patterns are able to be assist organizations with uncovering useful information then guide decision- making[9].

The Cross industry Standard Process for Data Mining (CRISP-DM) is a cycle process for development and analysis of data mining models [10]. As the demand for data mining increases and more algorithm are created, CRISP-DM ensures practices that everyone can follow, and it gives specific tips and techniques on how to understand business data by deploying a data- mining model. CRISP-DM has six phases including business understanding, data understanding, data preparation, modeling, evaluation, and deployment [10].

BACKRONG THEORY OF EDUCATIONAL DATA MINING

Educational data mining as a field solving educationally-related problems, at a high level, it seeks solutions to improve methods foe exploring the data, which usually has meaningful hierarchy at multiple levels, in order to discover new insights how people learn in the context of these setting [11]. For instances, a student's college transcript may contain a temporally ordered list of courses taken by him or her, the grade that the student earned in each course, and information about when the student selected or changed his or her academic major or minor. We might also understand how different individuals engage with or potentially 'game' the EDM system. Taken together, these learning analytics provide much useful information for the design of learning environments.

EDN applies los of techniques such as Decision Tree, Neural Networds, NaiveBayes, K-Nearest neighbor, etc into its examples. Qualitative techniques such as inter views and document analysis are frequently used to support case studies in EDM. EDM impact student directly within aeras of course content, course selections, recommender systems and admissions. In addition, application of specific data mining methods like web mining, classification and multivariate statistics are key techniques used in educationally related data [12]. These approaches can be applied into modeling students' individual difference and respond to those differences by providing a way, which help improve students' performance [13].

ALGORITHAMS OF DATA MINING

Data mining relies on disciplines like classification assists with identifying associations and clusters, and separates subjects under study. E.g, education institutions can use classification comprehensively to analyze student's characteristics. Categorization applies rule induction algorithms to handle categorical outcome variables. Estimation includes predictive functions or likelihood deals with continuous outcome variables. Estimation and classification use unsupervised or supervised modeling techniques. Visualization uses interactive graphs to demonstrate mathematically included data and scores, and is much more sophisticated than traditional bar charts or pie charts.

An algorithm is specifics, mathematically driven data mining function, such as a neural network, classification and regression tree (C&RT), or K-means. Data mining techniques including algorithms such as clustering, classification, regression, neural networks, association rules, decision tree, some of them have been applied successfully in the educational area [14]. E.g, methods for hierarchical

data mining and longitudinal data modeling have been applied into EDM.

Clustering

The goal of clustering is to find data points that naturally group together, splitting the full data set clusters sets. By using clustering techniques we are able to further identify dense and sparse regions in object space, and discover overall distribution pattern and correlations among data attributes. Here are some researches about clustering techniques [15-17, 18,19]

Classification

Classification is used to predict values for some variables. This algorithm frequently employs the decision tree or neural network-based classification algorithms. In classification test, data are used to estimate the accuracy of classification rules. If the accuracy is acceptable then rules are able to be apply into new data. Some popular classification methods include decision trees, logistic regression (for binary predications) and support vector machines.

Association rule

Association rules are used to find relations between different items[15]. Back to 1995, The analysis method of association rule was frequently utilize in most studies on educational data mining because of its less extensive expertise while comparing with other methods [20,21]. However, after the year of 2005, as researchers frequently adopting clustering and classification methods, the trend changed.

Regression

Regression analysis can be apply to model the relationship between independent variables and dependent variables. Independent variables are attributes we already known and response variables are able to predict what we want. But a number of real word problems are not simply prediction. Therefore, more complex techniques (e.g, decision tree, or neural nets) may be necessary for future prediction. Some popular regression methods within educational data mining include linear regression, neural networks, and support vector machine regression.

Neural network

Neural networks have the remarkable ability to obtain meaning from complicated or imprecise data and can be used extract patterns and detect trends that are too complex to be noticed by humans or computer techniques, they are good at identifying patterns or trends for future forecasting needs

Decision Trees

Decision tree is tree-shaped structures, which represents decisions sets. Specific decision tree

methods include classification and regression trees and Chi Squre Automatic Interaction Detection.

Nearest Neighbor Method
Nearest Neighbor Method, also called the k- nearest neighbor technique sometimes, classifies each record in the dataset based on a combination of different classes of the k record(s), which is similar to that in a historical dataset (where k >=1).
To choose the appropriate algorithms, researchers need design the data and align it with the desired output. As for small- scale, they can opt for clustering approach since it does not require necessary splitting data in classification.

METHOD OF EDUCATIONAL DATA MINING
EDM is a large number of powerful methods [2,22], some of which are widely acknowledged to be universal data mining types (E.g, clustering, prediction, outlier detecting, relationship mining, etc.). However, discovery with Models and Distillation of Data for Human judgment are considered more prominent methods recently in educational data mining[23].

Clustering
Clustering approaches had been applied to domain clear destination between the clusters. Once a set of clusters has been determined, new instances can be classified by determining the closet cluster. Clustering is able to be applied into grouping similar course materials or grouping students based on their learning behavior patterns [24]. There was an e-learning report was design to forecast the students' behavior pattern within the data mining approaches [25].

Prediction
Prediction can reduce single aspect of the data from combinations of data in other aspects. Classification, Regression and density estimation are main types of prediction methods. Prediction has already applied into predicting students' performance [26].

Relationship Mining
Relationship Mining can identify relationships among variables and encode them in rules for later application. Broadly, there are four different types of relationship mining: association rule mining, sequential pattern mining, correlation mining, and causal data mining, some of them has already utilized into identifying the relationship in patterns of students' behavior, difficulties or mistakes that learners usually encounter with [27].

Process Mining
Process mining has been used to extract data from event logs in an information system to from a clear

presentation in the overall activities. There are three different subfields of process mining: model discovery, model extension and conference checking. Process mining is reported able to reflect students' behavior in the sequence, grade etc.[28].

Text Mining
Text mining can be used for deriving information with high accuracy from text data and resources. The main contents of text mining contain text categorization, text clustering, concept/ entity extraction sentiment analysis and document summarization etc . Text mining is applied to analyze contents from forums, chats, Webpage or text resources etc.[29]

Social network analysis
Social network analysis (SNA) is used measure the relationship among different entities of information, and it is able to analyze the relationship in various tasks [30]

Discovery with Models
In discovery with models , a model is validated via prediction, clustering, or manual knowledge engineering. Key applications of this method include discovering relationships among student behaviors, characteristics and contextual variables [23, 31].

Distillation of data for Human Judgment
Data is distilled to enable humans to identify well-known patterns for identification, it is also distilled for classifying data features. The goal of this method is to summarize and present the information in a useful, interactive and visually appealing way for understanding the large amounts of education data and supporting decision making {32]. Data is distilled for human judgment in educational data mining for two main purpose: identification and classification.

APPLICATIONS IN EDUCATION
There are three applications of educational data mining with having received particular attention are discussed here.

Predicting Student performances
Lin applied classification and regression tree to predict what types of students would drop out from school, and then return to school later on [33]. The models were able to provide short-term accuracy for predicting which types of students would benefit from student retention programs [33]. Cha con and Spicer et al. developed a system based on data mining helps the institution identify and respond to students at-risk [34]. Their work is highly representative of the discipline, because it follows

with a strict data mining process, which quantitative. The research by Yeats, Reddy and Wheeler found that students who attend writing centers tend to do a good job in their classes [35]. Yu and DiGangi, et al. discovered that east coast students in USA tend to keep enrolled longer than their west coast counterparts do [36]. Course Management System EDM is often used in course management systems, like Moodle, which contains usage data that includes different activities. Garcia, Romero, Ventura, and de Castro developed a simplified data mining toolkit that operates within the course management system and allows students and their learning users to get data mining information their courses {37} This research and application contributions will allow non-technical faculty to engage in educational data mining activities. Instead of traditional static course patterns data mining can be applied to customize learning activities and adapt the pace for learners to complete courses [38,39]. It will create significant and optional learning experiences for each student. Also, Blikstein found different types of programming behaviors in an online course [40].

Planning and scheduling

Researches on mobile learning environments recently suggest that data mining can be applied to help provide personalized contents to different mobile user, despite the different between mobile devises and conventional PCs. EDM applications will allow non-technical users engage in data mining tools and activities making processing more accessible for all EDM user[39]. There are some examples, including statistical and visualization tools, analyzing social networks and related influence on learning outcomes [41].

CHALLENGES

As the related technology developed, costs and challenges associated with implementing EDM applications, like storing logged data and managing data systems [19]. Moreover, choosing which data to mine and analyze may also be a challenge. In addition, individual privacy is a continued concern for the application of educational data mining tools. With free, accessible tools in the market, students and learners may be at risk providing information to the learning system. Protecting individual privacy should be considered for the long-term development of EDM. Moreover, it's unclear what data displays, visualizations, and visual analysis are most informative and support effective decision making for different stakeholders

CONCLUSION AND FUTURE INSIGHTS

Data mining is powerful analytical tool to enhance decision making and analyzing new patterns and

relationships for organizations. And EDM contain techniques including data mining, statistic, machine learning, DM need to analyze data coming from teaching and learning, tests learning theories, and policy decision-making etc. There are a number of opportunities exist in EDM, from a analysis at organizational level to the analysis at individual level, even institutions.

Recently, there are several studies focus on applying EDM into admissions and enrollment, but we don't know exactly how institutions using data mining to enhance student learning or improving related educational processes. And results from EDM research are typically obtain from the narrow context specific educational settings. Therefore, the need for studies to examine in the border context is necessary. Furthermore, For the overall EDM work to be completed, the urgent need of examining how to widespread the adoption of educational data mining is concentrated in western cultures and subsequently, other countries like Asians may not be represented in the related researches and studies. Therefore, applications across multiple contexts should be considered in development of future models [39].

References

[1] Koedinger K, Cunningham K, SKogsholm A, Leber B. (2008) A n open repository and analysis tools for fine-grained, longitudinal learner data. In: First International Conference on Educational Data Mining. Montreal, Canada; 2008, 157-16.

[2] Baker, R.,& Yacef, K.(2009). The State of educational Data mining in 2009: A Review Future Visions. Journal of Educational Data Mining.1(1)

[3] Data Mining Curriculum. ACMSIGKDD.2006-04-30. Retrieved 2014-01-27.

[4] Clifton, Christopher (2010). Encyclopedia Britannica: Definition of Data Mining. Retrieved 2010-12-09.

[5] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009).The elements of statistical Learning: Data Mining, Inference, and Prediction. Retrieved 2012-08-07.specific educational settings. Therefore, the need for studies to examine in the border context is necessary. Furthermore, For the overall EDM work to be completed, the urgent need of examining how to widespread the adoption of educational data mining is concentrated in western cultures and subsequently, other countries like Asians may not be represented in the related researches and studies. Therefore, applications across multiple contexts should be considered in development of future models [39]. References

[1] Koedinger K, Cunningham K, SKogsholm A, Leber B. (2008) A n open repository and analysis tools for fine-grained, longitudinal learner data. In: First International Conference on Educational Data Mining. Montreal, Canada; 2008, 157-16.

[2] Baker, R.,& Yacef, K.(2009). The State of educational Data mining in 2009: A Review Future Visions. Journal of Educational Data Mining.1(1)

[3] Data Mining Curriculum. ACMSIGKDD.2006-04-30. Retrived 2014-01-27.

[4] Clifton, Christopher (2010). Encyclopedia Britannica: Definition of Data Mining. Retrieved 2010-12-09.

[5] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009).The elements of statistical Learning: Data Mining, Inference, and Prediction. Retrieved 2012-08-07. [6] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). From Data Mining to Knieledge Discovery in Databases. Retrieved 17 December 2008.

[7] Dunham, M. (2003). Data Mining: Introductory and Advanced Topic. Upper Saddle River, NJ: Person Education.

[8] Berson, A., Smith, S., & Thearling, K. (2011). An Overview Of Data Mining Techniques retrieved November 28, 2011, from

http://www.thearling.com/text/dmtechniques.htm

[9] Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2012). Analytics: The Widening Divide. MIT Sloan Management Review, 53(2), 1-22.

[10] 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,doi: 10.1057/dddmp. 2010.35.

[11] EducationalDataMining.org.(2013). Retrieved 2013-07-15.

[12] Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section oneducational data mining. SIGKDD Explor. Newsl., 13(2), 2-6. Doi:10.1145/2207243.2207245.

[13] Corbett, A. T. (2001). Cognitive Computer Tutors: Solving the Two-Sigma Problem. Paper presented at the Proceedings of the 8th International Conference on User Modeling 2001.

[14] Baker RSId (2010). Data mining for education. In MacGaw B, P. Baker E, eds. International Encyclopedia of education. 3rd ed. Vol.7.Oxford, UK: Elsevier; 2010,112-118.

[15] Pecheniz, M., Calders, T., Vasilyeva, E., & De Bra, P, (2008). Mining the student assessment data: Lessons drawn from a small scale case study. Educational Data Mining 2008, 187.

[16] He, W. (2012). Examining students' online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior.

[17] Shih, Y. C., Huang, P. R., Hsu, Y. C., & Chen, S. Y. (2012). A complete understanding of disorientation problems in web-based learning. The Turkish Online Journal of Educational Technology, 2012,11(3).

[18] Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar begavior groups in unstructured collaboration spaces. In Proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI 2004.

[19] Perera, D., Key, J., Koprinska, I., Yacef, K., & Zaiane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. Knowledge and Data Engineering, IEEE Transitions on, 21(6), 759-772.

[20] Romero, C., & Ventura, s. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135-146. [21] Merceron, A., & Yacef, K. (2007).Revisiting interestingness of strong symmetric association rules in educational data. In Proc. Of Int. Workshop on Applying Data Mining in e- Learning, Creete, Greece (pp. 3-12).

[22] Scheuer O, McLaren BM (2011). Educational data mining. In: The Encyclopedia of the Sciences of Learning. New York, NY: Springer; 2011.

[23] Baker, Ryan(2014). Data Mining for Education. Oxford, UK: Elsevier. Retrieved 9 Februray 2014.

[24] Vellido A, Castro F, Nebot A. (2011). Clustering Educational Data. Handbook of Educational Data Mianing. Boca Raton, FL: Chapman and Hall/CRC Press; 2011,75-92.

[25] Blagojevic, M., & Micic, Z. (2012). A web-based intelligent reporte-Learning system using data mining techniques. Computers & Electrical Engineering.

[26]Baker RSJd, Gowda SM, Corbett AT (2011). Automatically detecting a student's preparationfor future learning: help use is key. In:s. Eindhoven, The Netherlands; 2011, 179-188.

[27] Merceron A, Yacef K. (2010). Measuring correlation of strong symmetric association rules in Educational Data . In Romero C, Ventura S, Pechenizkiy M, Baker RSJd, eds. Handbook of Data Mining. Boca Raton, FL: CRC Press; 2010,245-256.

[28] TrckaN,PchenizkiyM, van derAalst W. (2011). Process mining from educational data. Hanbook of Educational Data Mining. Boca Raton, FL:CRC press; 2011, 123-142.

[29] Tane J, Schmitz C, Stumme G. (2004). Semantic resource management for the web: an e-learning application. In: International Conference Of the WWW. New York; 2004, 1-10.

[30] Rabbany R, Takaffoli M, Za"iane O. (2011). Analyzing participation of students in online cources using social network analysis techniques. In: International Coference on Educational Data Mining. Eindhoven, The Netherlands; 2011, 21-30.

[31] Biekowski M, Feng M, Means B. (2012). Enhancing teaching and learning through educational data mining and learning analytic: an issue brief. Washington, D.C: Office of Educational Technology, U.S. Department of Education; 2012, 1-57.

[32] Romero, Cristobal; Ventura, Sebastian (2013). WIREs Data Mining Knowl Discov. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3 (12): 12-27.Doi: 10.1002/widm.1075.

[33] Lin, S.-H. (2012). Data mining for student retention management. J Comput. Sci. Coll., 27(4), 92-99.

[34] Chacon, F., Spicer, D., & Valbuena, A. (2012). Analytic in support of student retention and Success (Research Bulletin 3, 2012 ed.). Louisville, CO: Educause center for app[35] Yeats, R., Reddy, P.J., Wheeler, A., Senior, C., & Murray, J. (2010). What a difference a writing centre makes: a small scale study. Education & Traning, 52(6/7), 499-507. Doi: 10.1108/00400911011068450.

[36] Yu, C. H., DiGangi, S., Jannasch-Pennell,A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of studentretention from sophomore to junior year. Journal of Data Science, 8, 307-325.

[37] Garcia, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. The Internet and Higher Education, 14(2), 77-88,doi:10.1016/j.iheduc.2010.07.006.

[38] Wang, Y.-h., & Liao, H.-C. (2011). Data mining for adaptive learning in a TESL learning system. Expert Systems with Applications, 38(6), 6480-6485. Doi: 10.1016/j.eswa.201011.098.

[39]Huebner, Richard A. (2014). A survey of educational data- mining research. Research in Higher Education Journal. Retrieved 30 March 2014.

[40] Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. Paper presented at the Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta,Canada.

[41] Handbook of educational data mining. CRC press. 2010.lied Research.