TEJAS Thiagarajar College Journal **ISSN(Online):2456-4044**
January 2018 Vol. 3(1) PP 41-46

Study on application of Clustering Algorithms in EDM

M. MUTHALAGU
PG Department of computer science
Thiagarajar college, Madurai.
Ph: 98420 62534 Email id:muthalagucs76@gmail.com

Abstract—Fifty years ago there were just a handful of universities across the globe that could provide for specialized educational courses. Today Universities are generating not only graduates but also massive amounts of data from their systems. So the question that arises is how can a higher educational institution harness the power of this didactic data for its strategic use? This review paper will serve to answer this question. To build an Information system that can learn from the data is a difficult task but it has been achieved successfully by using various data mining approaches like clustering, classification, prediction algorithms etc. However the use of these algorithms with educational dataset is quite low. This review paper focuses to consolidate the different types of clustering algorithms as applied in Educational Data mining context.

Keywords—clustering, educational data mining (EDM), learning styles, learning management systems (LMS)

I. INTRODUCTION

According to the international consortium on Educational Data Mining, EDM is defined 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 they learn in" [1].

EDM focuses on analyzing data generated in an educational setup by the various intra-connected or disparate systems to develop model for improving learning experience and institutional effectiveness. Data mining also sometimes referred to as knowledge discovery in databases (KDD) is a known field of study in life sciences and commerce but the application of data mining to educational context is limited [2].

Various methods have been proposed, applied and tested in data mining field and it's argued by some researchers that these generic methods or algorithms are not suitable to be applied to this emerging field of study.

It's proposed that educational data mining methods must be different from the standard data mining methods because of multi-level hierarchy and non-independence in educational data [1]. Institutions are increasing being held accountable for student success [3] since EDM emerged as a sub-discipline in DM there have been notable researches in student retention and attrition rates that have been conducted [4]. We applied predictive modeling technique to enhance student retention efforts. In a similar fashion, there have been various software's like Weka, Rapid Mineretc.that have been developed to use a combination of DM algorithms or a specific algorithm to aid researcher's or stakeholders to find

answers to specific problems but the problem with such tools are that they need to be learned so as to use them. This means that for a novice computer user especially in the administration department of a college or a university, the usage of such tools is not that easy. Just like commercial e-commerce based websites are using recommender systems that collect user browsing data and recommend similar products there have been efforts to apply the same in the educational context but they have not been successful as they are highly domain dependent [6].

The objective and purpose of this research paper is to review, different clustering algorithms as applied to EDM context. Numerous studies have been conducted in this context, but with disparate associations. This research paper is to bridge this gap and present a comprehensive review of all types of clustering methodologies as applied to EDM till date.

This paper is organized as follows. Section II is a background of related works pertaining to Educational Data Mining (EDM); Section III discusses the various clustering algorithms/techniques applied to educational dataset. Section IV discusses on the application of clustering algorithms to learning styles of student and learning management systems. Section V provides further discussion and finally Section VI shows the conclusion and future works.

II. EDUCATIONAL DATA MINING

"EDM converts raw data coming from educational systems into useful information. That could potentially have a greater impact on educational research and practice" [7]. Traditionally researchers have applied data mining methods like clustering, classification, association rule mining, text mining to educational context as outlined; [8], conducted a survey that provides a comprehensive resource of papers published between 1995 and 2005 on Educational Data Mining (EDM). Reference [9] has suggested the application of data mining techniques to study on-line courses. Had suggested association rules and clustering to support collaborative filtering for the development of more sensitive and effective e-learning systems. Reference [11] has used a case study that uses prediction methods in scientific study to game the interactive learning environment by exploiting the properties of the system rather than learning the system.

Reference [12] has provided tools that can be used to support educational data mining. Had shown how educational data mining prediction methods can be used to develop student models. It must be noted that student modeling is an emerging research discipline in educational data mining [1]. While another group of researchers have devised a toolkit that operates within the course management systems and is able to provide extracted mined information to non-expert users. Data mining techniques have been used to create dynamic learning exercises based on student's progress through a course on English language instruction. While most of the e-learning systems used by educational institutions are used to post or access course materials, they do not provide the educators the necessary tools that could thoroughly track and evaluate all the activities performed by their learners so as to evaluate the effectiveness of the course and learning process.

III. CLUSTERING TECHNIQUES

The theory of looking at didactic amounts of data whether it's in digital or physical form and stored in diverse repositories be it book keeping records or databases of an educational institution is now termed as big data. According, to Manyika a data set whose computational size exceeds the processing limit of software can be categorized as big data. Several studies have been conducted in the past that have provided detailed insights into the application of traditional data mining algorithms like clustering, prediction, association to tame the sheer voluminous power of big data [9]. Traditional Data Mining algorithms have been applied to various kinds of educational systems. Broadly, the educational system can be classified as two types, brick and mortar based traditional classroom's and the digital virtual classroom's better known as known as LMS Systems , web-based adaptive hypermedia systems and intelligent tutoring systems (ITS). The application of various clustering algorithm has been applied in many a cases to educational data set in diverse studies. The following table consolidates the research work done on the application of clustering algorithms to educational dataset.

IV. USING CLUSTERING IN EDM

In a learning environment the learning styles of student is a decisive factor. In many cases there has been a mismatch between personal learning styles and the learning demands of different disciplines. Reference, has utilized a two step cluster analysis approach which examined the brain signals centroids that used electroencephalography (EEG) technology to measure the learning style of participants such that they were successfully able to classify it into 4 unique clusters. Students typically annotate texts while reading book by highlighting the context of interest or by underlining it or by writing comments in the side margins. This activity is called annotation. Researchers have applied statistical clustering method like K-means clustering and Hierarchical clustering to student annotations. And they proved that by using these clustering methods, the creation of students with similar learning style cluster is improved and is faster. Comprehension reading is a very widely used classroom activity in schools and colleges. This helps in building a lifelong reading habit and learning process. This ability of the student behavioral learning patterns has been computationally mapped by applying the Forgy method for k-means clustering and combined with Bloom's taxonomy to determine positive and negative cognitive skills set in reference to reading comprehension skills. Yet in another study, combined Web Based Instruction (WBI) programs with the cognitive learning style of the learner to study their effects on student learning patterns. K-means clustering algorithm was used to result in cluster of students that shared similar learning patterns that further leads to identification of the related cognitive style for each group.

Learning Management System (LMS) have become an integral part of educational institutions for teaching and learning. A typical LMS logs most of the user activities like course attempted, modules read, practice exam attempted, exam score, student-

student interaction via chat logs or discussion boards similarly student-teacher interaction via discussion boards is also logged in the LMS. Several studies have been conducted in this regard.

Reference studied the usage statistics that an LMS provides and worked on its statistical data analysis and the results were applied in the University of Valencia (Spain) .Although they were successful in the statistical analysis of LMS usage data using SPSS but to standardize their methodology the subsequent automation process is yet to be completed and has been left as a future work. Performance in exams, usage statistics, regression, number of visits, top search terms, number of downloads of e-learning resources is presented .Several DM approaches and techniques (clustering, classification and association analysis) have been proposed for joint use in the mining of student's assessment data in LMS. Association rules, clustering, classification, sequential pattern analysis, dependency modeling, and prediction have been used to improve web based learning environments to subsequently enhance the degree to which the educator can evaluate the learning process. Analysis of user access log in Moodle to improve e-learning and to support the analysis of trends is presented in Comparison of different DM algorithms are made to classify learners (predict final marks) based on Moodle usage data .Prediction of student's performance (final grade) based on features extracted from logged data is presented in and university academic student performance is presented. Prediction of online student's marks (using an orthogonal search-based rule extraction algorithm) is presented.

Other studies have been conducted to predict student's performance from log and test scores in web-based instruction (using multi-variable regression), While have used classification, clustering, association rule mining and regression for the discovery of possible dependencies among learner's mean performance and course characteristics. Their results confirm that students behavior in an online learning platform affect their performance.

In another study, researchers have shown how educational institutions can benefit from the data collected by LMS. They have proposed an algorithm called "Course Classification Algorithm" when applied in the LMS (Open e-Class platform) that the institution uses can be used to determine and generate course content quality and student online usage reports. These reports are then sent to the instructors for evaluation and motivation purpose. Have proposed the usage of k-means clustering and self organizing map to cluster learning objects (learning objects are educational resources like eBook, question paper, answer index etc.) so as to facilitate faster accessibility of such resources by searching in a LMS. Have proposed Particle Swarm Optimization (PSO)-based clustering for improving the quality of learning by integrating Personalized Learning Environment (PLE) in conjunction with the conventional Learning Management System (LMS)

However, one of the major problems that researchers encounter in finding interesting patterns from educational data set is the relatively small size of the dataset. In another study, we have applied Expectation- Maximization (EM) clustering

algorithm to discover student profiles from course evaluation data and for finding associations between subjects that was based on student performance.

Employability of its graduates has been a primary goal of higher educational institutions. Knowledge workers are resorting to key educational courses using Massive Open Online Courses (MOOCs) being provided online by institutions of repute like MIT, Stanford, Harvard to name a few. The year 2012 was witness to a rapid development and expansion of several MOOEPs (Massive Open Online Education Platform) like Canvas, Class Togo, Coursera,edX, NPTEL, Udacity to name a few. The service provider / knowledge workers had conducted study to explore the scope of interdisciplinary education through MOOCs. Employability education is an integral component of higher education and an important path by which companies obtain excellent employees. It has been a sustainable argument that in the present socio- economic development, the employability based educational content becomes a mandate.

V. DISCUSSION

So far we see that subject specific research has been done but what about domain specific i.e. how do institutions employ or apply data mining methods to improve on institutional effectiveness? Zimmerman's educational model states that maintaining and monitoring student's academic record is an integral activity of an educational institution. The researchers had used classification algorithm and Prediction algorithm namely decision table and One R algorithm on students" academic record from a previous semester to predict their performance in the current semester. An educational institution maintains and stores various types of student data, it can range from student academic data to their personal record like parents income, parent's qualification etc. In a study conducted by they have proved that student's performance can be predicted by using a data set that consisted of student's gender, its parental education, its financial background etc. The researchers have used Bayesian networks to predict the student outcome based on attributes like attendance, performance in class tests, assignments etc. Researchers have applied data mining methods like dimensional modeling into educational institutions while others like have used regression analysis and classification (CS5.0 algorithm which is a type of decision tree) to predict the academic dismissal of students and to predict the GPA of graduated students in e-learning center.

VI. CONCLUSION AND FUTURE WORK

The application of data mining methods in the educational sector is an interesting phenomenon. It sets to uncover the previously hidden data to meaningful information that could be used for both strategic as well as learning gains. In this review paper, we have detailed the various disparate entities that are widely spread across in the educational foray. However, collectively they have not been addressed and this paper serves to bridge this gap. We would continue to pursue our research in clustering algorithms as applied to educational context and will also be working towards generating a unified clustering approach such that it could easily be

applied to any educational institutional dataset without any much overhead.

REFERENCES

[1] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," Journal of Educational Data Mining, vol. 1, no. 1, 2009.

[2] J. Ranjan and K. Malik, "Effective educational process: A datamining approach," Vine, vol. 37, no. 4, pp. 502-515, 2007.

[3] B. J. P. Campbell, P. B. Deblois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," Educause Review, vol. 42, pp. 40-57, 2007.

[4] J. Luan, Data Mining and Knowledge Management in Higher Education, Toronto, Canada, 2002.

[5] S. Lin, "Data mining for student retention management," J. Comput. Sci. Coll., vol. 27, no. 4, pp. 92-99, 2012.

[6] O. C. Santos and J. G. Boticario, "Modeling recommendations for the educational domain," Procedia Comput. Sci., vol. 1, no. 2, pp.
2793-2800, Jan. 2010.

[7] Yan Yan, Qin Xingbin. A Review of Big Data Research in Medicine & Healthcare. e-Science Technology & Application, 2014, 5(6): 3-16.

[8] De Oliveira M C F, Levkowitz H. From visual data exploration to visual data mining: a survey [J]. IEEE Transactions on Visualization and Computer Graphics, 2003, 9(3): 378-394.

[9] J. R. QUINLAN. Programming for Machine Learning [M]. CA: San Mateo, 1993.

[10] Simeon J. Michael H. Arturas Mazeika. Visual Data Mining □An Introduction and Overview□ Springer Berlin/Heidelberg□2008.

[11] Bertini E, Lalanne D. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery[C] //Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. New York: ACM Press, 2009: 12-20.

[12] J.A. Fails and J. Olsen, "Interactive machine learning," IUI'03: Proceedings of the 8th international conference on intelligent user interfaces, New York, NY, USA: ACM, 2003, pp. 39-45.