International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

Big Data in Educational Data Mining and Learning Analytics

[1] D.Varun [2] Arun Prasad Desai [3] G. N. S. Pravallika
[1][3]B. Tech (CSE) [2] Assistant Professor
[1][2][3] Department of CSE Sri Venkateswara College of Engineering, Nellore

***Abstract*—Educational data mining and learning analytics are used to research and build models in several areas that can influence learning systems. Higher education institutions are beginning to use analytics for improving the services they provide and for increasing student grades and retention. With analytics and data mining experiments in education starting to proliferate, sorting out fact from fiction and identifying research possibilities and practical applications are not easy. This issue brief is intended to help policymakers and administrators understand how analytics and data mining have been—and can be—applied for educational improvement. At present, educational data mining tends to focus on developing new tools for discovering patterns in data. These patterns are generally about the micro concepts involved in learning, learning analytics**

***Keywords:*--Data mining, big data, and Educational data mining, learning, learning analytics**

I. INTRODUCTION

Research on machine learning has yielded techniques for knowledge discovery or data mining that discover novel and potentially useful information in large amounts of unstructured data. These techniques find patterns in data and then build predictive models that probabilistically predict an outcome. Applications of these models can then be used in computing analytics over large datasets. Two areas that are specific to the use of big data in education are educational data mining and learning analytics. Although there is no hard and fast distinction between these two fields, they have had somewhat different research histories and are developing as distinct research areas. Generally, educational data mining looks for new patterns in data and develops new algorithms and/or new models, while learning analytics applies known predictive models in instructional systems.

II. EDUCATIONAL DATA MINING

Educational data mining is emerging as a research area with a suite of computational and psychological methods and research approaches for understanding how students learn [1]. New computer-supported interactive learning methods and tools—intelligent tutoring systems, simulations, games—have opened up opportunities to collect and analyze student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn. Data collected from online learning systems can be aggregated over large numbers of students and can contain many variables that data mining algorithms can explore for model building.

Just as with early efforts to understand online behaviors, early efforts at educational data mining involved mining website log data [2], but now more integrated, instrumented, and sophisticated online learning systems provide more kinds of data. Educational data mining generally emphasizes reducing learning into small components that can be analyzed and then influenced by software that adapts to the student [3]. Student learning data collected by online learning systems are being explored to develop predictive models by applying educational data mining methods that classify data or find relationships. These models play a key role in building adaptive learning systems in which adaptations or interventions based on the model's predictions can be used to change what students experience next or even to recommend outside academic services to support their learning.

An important and unique feature of educational data is that they are hierarchical. Data at the keystroke level, the answer level, the session level, the student level, the classroom level, the teacher level, and the school level are nested inside one another [4, 5]. Other important features are time, sequence, and context. Time is important to capture data, such as length of practice sessions or time to learn. Sequence represents how concepts build on one another and how practice and tutoring should be ordered. Context is important for explaining results and knowing where a model may or may not work. Methods for hierarchical data mining and longitudinal data modeling have been important developments in mining educational data

Educational data mining researchers [6] view the following as the goals for their research:

2. Predicting students' future learning behavior by creating student models that incorporate such detailed information as students' knowledge, motivation, metacognition, and attitudes;

3. Discovering or improving domain models that characterize the content to be learned and optimal instructional sequences;

4. Studying the effects of different kinds of pedagogical support that can be provided by learning software; and

5. Advancing scientific knowledge about learning and learners through building computational models that incorporate models of the student, the domain, and the software's pedagogy.

To accomplish these four goals, educational data mining research uses the five categories of technical methods [7] described below.

❖ ***Prediction*** entails developing a model that can infer a single aspect of the data (predicted variable) from somecombination of other aspects of the data (predictor variables). Examples of using prediction include detecting such student behaviors as when they are gaming the system, engaging in off-task behavior, or failing to answer a question correctly despite having a skill. Predictive models have been used for understanding what behaviors in an online learning environment—participation in discussion forums, taking practice tests and the like—will predict which students might fail a class. Prediction shows promise in developing domain models, such as connecting procedures or facts with the specific sequence and amount of practice items that best teach them, and forecasting and understanding student educational outcomes, such as success on posttests after tutoring [8].

❖ ***Clustering*** refers to finding data points that naturally group together and can be used to split a full dataset intocategories. Examples of clustering applications are grouping students based on their learning difficulties and interaction patterns, such as how and how much they use tools in a learning management system[9], and grouping users for purposes of recommending actions and resources to similar users. Data as varied as online learning resources, student cognitive interviews, and postings in discussion forums can be analyzed usingtechniques for working with unstructured data to extract characteristics of the data and then clustering the results. Clustering can be used in any domain that involves classifying, even to determine how much collaboration users exhibit based on postings in discussion forums[10]

❖ ***Relationship mining*** involves discovering relationships between variables in a dataset and encoding them as rulesfor later use. For example, relationship mining can identify the relationships among products purchased in online shopping [11].

❖ ***Association rule mining*** can be used for finding student mistakes that co-occur, associating content with user types tobuild recommendations for content that is likely to be interesting, or for making changes to teaching approaches [12]. These techniques can be used to associate student activity, in a learning management system or discussion forums, with student grades or to investigate such questions as why students' use of practice tests decreases over a semester of study.

❖ ***Sequential pattern mining*** builds rules that capture the connections between occurrences of sequential events, forexample, finding temporal sequences, such as student mistakes followed by help seeking. This could be used to detect events, such as students regressing to making errors in mechanics when they are writing with more complex and critical thinking techniques, and to analyze interactions in online discussion forums.

Key educational applications of relationship mining include discovery of associations between student performance and course sequences and discovering which pedagogical strategies lead to more effective or robust learning. This latter area—called teaching analytics—is of growing importance and is intended to help researchers build automated systems that model how effective teachers operate by mining their use of educational systems.

❖ ***Distillation*** for human judgment is a technique that involves depicting data in a way that enables a human to quicklyidentify or classify features of the data. This area of educational data mining improves machine-learning models because humans can identify patterns in, or features of, student learning actions, student behaviors, or data involving collaboration among students. This approach overlaps with visual data analytics.

❖ ***Discovery with models*** is a technique that involves using a validated model of a phenomenon (developed throughprediction, clustering, or manual knowledge engineering) as a component in further analysis. A sample student activity discerned from the data was —map probing.‖ A model of map probing then was used within a second model of learning strategies and helped researchers study how the strategy varied across different experimental states. Discovery with models supports discovery of relationships between student behaviors and student characteristics or contextual variables, analysis of research questions across a wide variety of contexts, and integration of psychometric modeling frameworks into machine-learned models.

❖ ***Prediction:*** entails developing a model that can infer a single aspect of the data (predicted variable) from somecombination of other aspects of the data (predictor variables). Examples of using prediction include detecting such student behaviors as when they are gaming the system, engaging in off-task behavior, or failing to answer a question correctly despite having a skill. Predictive models have been used for understanding what behaviors in an online learning environment—participation in discussion forums, taking practice tests and the like—will predict which students might fail a class. Prediction shows promise in developing domain models, such as connecting procedures or facts with the specific sequence and amount of practice items that best teach them, and forecasting and understanding student educational outcomes, such as success on posttests after tutoring [8].

❖ ***Clustering*** refers to finding data points that naturally group together and can be used to split a full dataset intocategories. Examples of clustering applications are grouping students based on their learning difficulties and interaction patterns, such as how and how much they use tools in a learning management system [9], and grouping users for purposes of recommending actions and resources to similar users. Data as varied as online learning resources, student cognitive interviews, and postings in discussion forums can be analyzed using techniques for working with unstructured data to extract characteristics of the data and then clustering the results. Clustering can be used in any domain that involves classifying, even to determine how much collaboration users exhibit based on postings in discussion forums[10]

❖ ***Relationship mining*** involves discovering relationships between variables in a dataset and encoding them as rulesfor later use. For example, relationship mining can identify the relationships among products purchased in online shopping [11].

❖ ***Association rule mining*** can be used for finding student mistakes that co-occur, associating content with user types tobuild recommendations for content that is likely to be interesting, or for making changes to teaching approaches [12]. These techniques can be used to associate student activity, in a learning management system or discussion forums, with student grades or to investigate such questions as why students' use of practice tests decreases over a semester of study.

❖ ***Sequential pattern mining*** builds rules that capture the connections between occurrences of sequential events, forexample, finding temporal sequences, such as student mistakes followed by help seeking. This could be used to detect events, such as students regressing to making errors in mechanics when they are writing with more complex and critical thinking techniques, and to analyze interactions in online discussion forums.

Key educational applications of relationship mining include discovery of associations between student performance and course sequences and discovering which pedagogical strategies lead to more effective or robust learning. This latter area—called teaching analytics—is of growing importance and is intended to help researchers build automated systems that model how effective teachers operate by mining their use of educational systems.

❖ ***Distillation*** for human judgment is a technique that involves depicting data in a way that enables a human to quicklyidentify or classify features of the data. This area of educational data mining improves machine-learning models because humans can identify patterns in, or features of, student learning actions, student behaviors, or data involving collaboration among students. This approach overlaps with visual data analytics.

❖ ***Discovery with models*** is a technique that involves using a validated model of a phenomenon (developed throughprediction, clustering, or manual knowledge engineering) as a component in further analysis. A sample student activity discerned from the data was —map probing.‖ a model of map probing then was used within a second model of learning strategies and helped researchers study how the strategy varied across different experimental states. Discovery with models supports discovery of relationships between student behaviors and student characteristics or contextual variables, analysis of research questions across a wide variety of contexts, and integration of psychometric modeling frameworks into machine-learned models.

III. LEARNING ANALYTICS

Learning analytics is becoming defined as an area of research and application and is related to academic analytics, action analytics, and predictive analytics. Learning analytics draws on a broader array of academic disciplines than educational data mining, incorporating concepts and techniques from information science and sociology, in addition to computer science, statistics, psychology, and the learning sciences. Unlike educational data mining, learning analytics generally does not emphasize reducing learning into components but instead seeks to understand entire systems and to support human decision making. Learning analytics emphasizes measurement and data collection as activities that institutions need to undertake and understand, and focuses on the analysis and reporting of the data. Unlike educational data mining, learning analytics does not generally address the development of new computational methods for data analysis but instead addresses the application of known methods and models to answer important questions that affect student learning and organizational learning systems. Unlike educational data mining, which emphasizes system-generated and automated responses to students, learning analytics enables human tailoring of responses, such as through adapting instructional content, intervening with at-risk students, and providing feedback.

Technical methods used in learning analytics are varied and draw from those used in educational data mining. Additionally, learning analytics may employ:

❖ Social network analysis (e.g., analysis of student-to-student and student-to-teacher relationships and interactions to identify disconnected students, influencers, etc.) and

❖ Social or ―attention‖ metadata to determine what a user is engaged with.

As with educational data mining, providing a visual representation of analytics is critical to generate actionable analyses; information is often represented as ―dashboards‖ that show data in an easily digestible form.

A key application of learning analytics is monitoring and predicting students' learning performance and spotting potential issues early so that interventions can be provided to identify students at risk of failing a course or program of study [13]. Several learning analytics models have been developed to identify student risk level in real time to increase the students' likelihood of success. Examples of such systems include Purdue University's Course Signals system and the Moodog system being used at the course level at the University of California, Santa Barbara, and at the institutional level at the University of Alabama. Higher education institutions have shown increased interest in learning analytics as they face calls for more transparency and greater scrutiny of their student recruitment and retention practices.

Data mining of student behavior in online courses has revealed differences between successful and unsuccessful students in terms of such variables as level of participation in discussion boards, number of emails sent, and number of quizzes completed. Analytics based on these student behavior variables can be used in feedback loops to provide more fluid and flexible curricula and to support immediate course alterations based on analyses of real-time learning data.

IV.EDUCATIONAL DATA MINING AND LEARNING ANALYTICS APPLICATIONS

Educational data mining and learning analytics research are beginning to answer increasingly complex questions about what a student knows and whether a student is engaged. For example, questions may concern what a short-term boost in performance in reading a word says about overall learning of that word, and whether gaze-tracking machinery can learn to detect student engagement. Researchers have experimented with new techniques for model building and also with new kinds of learning system data that have shown promise for predicting student outcomes. Previous sections presented the research goals and techniques used for educational data mining and learning/visual analytics. This section presents broad areas of applications that are found in practice, especially in emerging companies. These application areas were discerned from the review of the published and gray literature and were used to frame the interviews with industry experts. These areas represent the broad categories in which data mining and analytics can be applied to online activity, especially as it relates to learning online. This is in contrast to the more general areas for big data use, such as health care, manufacturing, and retail [14].

These application areas are (1) modeling of user knowledge, user behavior, and user experience; (2) user profiling; (3) modeling of key concepts in a domain and modeling a domain's knowledge components, (4) and trend analysis. Another application area concerns how analytics are used to adapt to or personalize the user's experience. Each of these application areas uses different sources of data, and Exhibit, briefly describes questions that these categories answer and lists data sources that have been used thus far in these applications. In the remainder of this section, each area is explored in more detail along with examples from both industry practice and academic research.

V.CONCLUSION

Working with big data using data mining and analytics is rapidly becoming common in the commercial sector. Tools and techniques once confined to research laboratories are being adopted by forward-looking industries, most notably those serving end users through online systems. Higher education institutions are applying learning analytics to improve the services they provide and to improve visible and measurable targets such as grades and retention. Now, with advances in adaptive learning systems, possibilities exist to harness the power of feedback loops at the level of individual teachers and students. Measuring and making visible students' learning and assessment activities open up the possibility for students to develop skills in monitoring their own learning and to see directly how their effort improves their success. Teachers gain views into students' performance that help them adapt their teaching or initiate interventions in the form of tutoring, tailored assignments, and the like. Adaptive learning systems enable educators to quickly see the effectiveness of their adaptations and interventions, providing feedback for continuous improvement. Researchers and developers can more rapidly compare versions of designs, products, and approaches to teaching and learning, enabling the state of the art and the state of the practice to keep pace with the rapid pace of adoption of online and blended learning environments.

REFERENCES

1) Anaya, A. R., and J. G. Boticario. 2009. ―A Data Mining Approach to Reveal Representative Collaboration Indicators in Open Collaboration Frameworks.‖ In Educational Data Mining 2009: Proceedings of the 2nd International Conference on Educational Data Mining, edited by T. Barnes, M. Desmarais, C. Romero, and S. Ventura, 210–219.

2) Amershi, S., and C. Conati. 2009. ―Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments.‖ Journal of Educational Data Mining 1 (1): 18–71.

3) Arnold, K. E. 2010. ―Signals: Applying Academic Analytics. EDUCAUSE Quarterly 33 (1). http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/SignalsApplyingAcademicAnalyti/199385

4) Baker, R. S. J. d., S. M. Gowda, and A. T. Corbett. 2011. ―Automatically Detecting a Student's Preparation for Future Learnin g: Help Use Is Key.‖ In Proceedings of the 4th International Conference on Educational Data Mining, edited by M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. Stamper, 179–188.

5) Blikstein, P. 2011. ―Using Learning Analytics to Assess Students' Behavior in Open-Ended Programming Tasks.‖ Proceedings of the First International Conference on Learning Analytics and Knowledge. New York, NY: Association for Computing Machinery, 110–116.

6) Jeong, H., and G. Biswas. 2008. ―Mining Student Behavior Models in Learning-by-Teaching Environments.‖ In Proceedings of the 1st International Conference on Educational Data Mining, Montréal, Québec, Canada,127–136.

7) Köck, M., and A. Paramythis. 2011. ―Activity Sequence Modeling and Dynamic Clustering for Personalized E-Learning. Journal of User Modeling and User-Adapted Interaction 21 (1-2): 51–97.

8) Koedinger, K. R., R. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. 2010. ―A Data Repository for the EDM Community: The PSLC DataShop.‖ In Handbook of Educational Data Mining, edited by C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker. Boca Raton, FL: CRC Press, 43–55.

9) Long, P. and Siemens, G. 2011. ―Penetrating the Fog: Analytics in Learning and Education.‖ EDUCAUSE Review 46 (5). New Media Consortium. 2012. NMC Horizon Project Higher Ed Short List. Austin, TX: New MediaConsortium. http://www.nmc.org/news/download-horizon-project-2012-higher-ed-short-list.

10) Siemens, G., and R. S. J. d. Baker. 2012. ―Learning Analytics and Educational Data Mining: Towards Communication and Collaboration.‖ In Proceedings of LAK12: 2nd International Conference on Learning Analytics & Knowledge, New York, NY: Association for Computing Machinery, 252–254.

11) Lauría, E. J. M., and J. Baron. 2011. Mining Sakai to Measure Student Performance: Opportunities and Challenges in Academic Analytics.http://ecc.marist.edu/conf2011 /materials /LauriaECC2011-%20Mining%20Sakai%20to%20 Measure % 20Student%20Performance%20-%20final.pdf

12) Romero C. R. and S. Ventura. 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.