Using Data Mining for Improving Education of University Students – A Survey

K. P. S. D. Kumarapathirana
Sri lanka

Abstract

Data mining combines machine learning, statistical and visualization techniques to discover and extract knowledge. Student retention is an indicator of academic performance and enrolment management of the university. Poor student retention could reflect badly on the university. Universities are facing the immense and quick growth of the volume of educational data stored in different types of databases and system logs. Moreover, the academic success of students is another major issue for the management in all professional institutes. So the early prediction to improve the student performance through counseling and extra coaching will help the management to take timely action for decrease the percentage of poor performance by the students. Data mining can be used to find relationships and patterns that exist but are hidden among the vast amount of educational data. This survey conducts a literature survey to identify data mining technologies to monitor student, analyze student academic behavior and provide a basis for efficient intervention strategies. The results can be used to develop a decision support system and help the authorities to timely actions on weak students.

Introduction

Simply, data mining is considered as a process that takes data as input and outputs knowledge which initially was known as the Knowledge Discovery in Databases (KDD) process. Many other terms are being used to interpret data mining, such as knowledge mining from databases, knowledge extraction, data analysis, and data archaeology. It is also defined as a nontrivial process of identifying valid, novel, potentially useful, and ultimately logical patterns in data (U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 1996). This has been aided by some other techniques in the field of computer science, such as neural networks, classification, clustering, genetic algorithms, association rules and support vector machines. Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns.
Data mining is an emerging powerful tool for analysis and prediction. It is successfully applied in different areas such as fraud detection, advertising, marketing, loan assessment and prediction. But, it is in emerging stage in the field of education. This research would focus on a comprehensive survey, a travelogue (2000-2018) towards educational data mining and its scope in future.
One of the primary goals of the educational system any higher education institute must focus on preparing students with the knowledge and skills needed to convert into successful careers within a specified period (especially within three or four years). How effectively these educational systems meet this goal is a major determinant of both economic and social progress of any country.

There are currently fifteen universities in Sri Lanka, which are established under the authority of the University Grants Commission (Universities, 2018). All these universities accept approximately 30,000 students in each year. It has become difficult to provide high quality teaching and guidance to such a large number of students. As a result, many students fail to complete their degrees within the required periods. Using data mining (DM) techniques to analyze student information can help identify possible reasons for student failures.

Within recent few years, the number of educational institutes that adopted an information system has been growing very quickly; consecutively the amount of data available in each educational institute database has also increased. Educational data mining is intuitively applied to discover hidden information from this data that would improve the quality of the whole educational system. Educational data mining can be applied to discover patterns in untrusted datasets to automate the decision making process of learners, students and administrators.

This knowledge extracted from students' data can be used in different ways such as to validate and evaluate an educational system, improve the quality of T& L processes, and lay the groundwork for a more effective learning process (C. Romero, S. Ventura, and P. De Bra, 2004). Similar ideas have been applied successfully, especially in business data, in different datasets, such as e-commerce systems, to increase sales profits.

If universities could identify the factors which affect the low performance as earlier as possible and hence are able to predict students' behavior in different perspectives, the administration can use this knowledge in taking actions in order to overcome issues related and to improve the performance of such students. It will be a win-win situation for all the stakeholders of universities/institutions i.e. management, teachers, students and parents. Students will be able to identify their weaknesses beforehand and can improve themselves. Teachers will be able to plan their lectures as per the need of students and can provide better guidance to such students. Parents will be reassured of their ward performance in such institutes. Management can bring in better policies and strategies to enhance the performance of these students with additional facilities. Eventually, this will help in producing skillful workforce and hence sustainable growth for the country.

Literature Survey

Castro (2007) categorized data mining tasks used in higher educational institutes into four different areas: applications that deal with the assessment of students learning performance, course adaptation and learning recommendations to customize students learning based on individual students behaviors, developing a method to evaluate materials in online courses, approaches that use feedback from students and teachers in e-learning courses, and detection models for uncovering student learning behaviors. Later, Baker and Yacef (2009) suggested four key areas of educational data mining application, namely, improving student models, improving domain models, studying the pedagogical support provided by learning software, and conducting scientific research on learning and learners using five

approaches/methods: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models.

Work related to data mining process can be divided into two main categories: data mining and visualization. The category of statistics and visualization has received a prominent place in theoretical discussions and research (C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker, 2010) (R. S. Baker and K. Yacef, 2009). Baker (R. S. Baker and K. Yacef, 2009) classifies the work as follows:

1. Prediction.
    - Classification.
    - Regression.
    - Density estimation.
2. Clustering.
3. Relationship mining.
    - Association rule mining.
    - Correlation mining.
    - Sequential pattern mining.
    - Causal DM.
4. Distillation of data for human judgment.
5. Discovery with models.

Discovery with models category identifies which learning material subcategories provide students with the most benefits (J. Mostow and J. Beck, 2008), how specific students' behavior affects students learning in different ways (M. Cocea, A. Hershkovitz, and R. S. Baker, 2009), and how tutorial design affects students learning (H. Jeong and G. Biswas, 2008).

Outlier detection is another educational data mining methodology which has not been used widely. It discovers data points that significantly differ from the rest of the data (V. J. Hodge and J. Austin, 2014). In educational data mining, they can detect students with learning problems and irregular learning practices by using the learners' response time data (Chan, 2007) and deviations in teaching learning activities (Muehlenbrock, 2005).

Text mining which works with datasets such as text documents, HTML files, emails, etc., has been used in this area to analyze data in Learning Management Systems (Ueno, 2004), (L. P. Dringus and T. Ellis, 2005) and in web content mining (J. Chen, Q. Li, L. Wang, and W. Jia, 2004). Use of text mining for the clustering of documents based on similarity and topic has been proposed (J. Tane, C. Schmitz, and G. Stumme, 2004), (C. Tang, R. W. Lau, Q. Li, H. Yin, T. Li, and D. Kilis, 2000).

Prediction methodology studies features used for prediction and uses those features in the underlying construct to predict student educational outcomes (C. Romero, S. Ventura, P. G. Espejo, and C. Herv´as, 2008).

Association rule mining is the most common educational data mining method. The relationship found in association rule mining is {if –> then} rules. For example, if {Student GPA is less than two, and the student has a job} → {the student is going to drop out of school}. The main goal of relationship mining is to

determine whether or not one event causes another event by studying the coverage of the two events in the data set (C. Wallace, K. B. Korb, and H. Dai, 1996), or by studying how an event is triggered.

Data used in Mining Educational Data

Different data mining techniques are used to analyze educational data and solve issues related to education. Similar to other data mining techniques, it is needed to extract interesting, interpretable, useful, and novel information from educational data. However, this is specifically concerned with developing methods to explore the unique types of data in educational settings (S. K. Mohamad and Z. Tasir, 2013). Offline education, also known as traditional education, is where knowledge transfers to learners based on face-to-face contact.

Data which can be collected by traditional methods such as observation and questionnaires is one of the source of data used in educational data mining. It studies the cognitive skills of students and determines how they learn. Therefore, the statistical technique and psychometrics can be applied to the data.

Another way to collect data is from materials, instruction, communication, and reporting tools that allow them to learn by themselves used in e-learning and learning management systems (LMS). Data mining techniques can be applied to the data stored by the systems in the databases.

Intelligent tutoring systems (ITS) and adaptive educational hypermedia systems (AEHS) try to customize the data provided to students based on student profiles. As a result, applying data mining techniques is important for building user profiles.

Based on the above sources, we can group EDM research according to the type of data used: traditional education, web-based education (e-learning), learning management systems, intelligent tutoring systems, adaptive educational systems, tests questionnaires, texts contents, and others.

Summary

Prediction develops a model to predict some variables base on other variables. The predictor variables can be constant or extract from the data set. This identifies at-risk students and understand student educational outcomes.

Clustering groups specific amount of data to different clusters based on the characteristics of the data. The number of clusters can be different based on the model and the objectives of the clustering process. This finds similarities and differences between students or schools and categorizes new student behavior.

Relationship Mining extracts the relationship between two or more variables in the data set. This finds the relationship between parent education level and students drooping out from school. Discovery of curricular associations in course sequences and discovering which pedagogical strategies lead to more effective/robust learning.

Moreover, some models aim to develop a model of a phenomenon using clustering, prediction, or knowledge engineering, as a component in more comprehensive model of prediction or relationship

mining. Discovery of relationships between student behaviors, and student characteristics or contextual variables; Analysis of research question across wide variety of contexts

With the ability to uncover hidden patterns in large databases, universities can build models that predict with a high degree of accuracy, the behavior of population clusters. By acting on these predictive models, educational institutions can effectively address issues ranging from transfers and retention, to marketing and alumni relations.

References

C. Romero, S. Ventura, and P. De Bra. (2004). Knowledge discovery with genetic programming for providing feedback to courseware authors. *User Modeling and User-Adapted Interaction, 14*(5), 425–464.

C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker. (2010). *Handbook of educational data mining.* CRC Press.

C. Romero, S. Ventura, P. G. Espejo, and C. Herv´as. (2008). Data mining algorithms to classify students. *EDM*, 8-17.

C. Tang, R. W. Lau, Q. Li, H. Yin, T. Li, and D. Kilis. (2000). Personalized courseware construction based on web data mining. *Web Information Systems Engineering, 2000. Proceedings of the First International Conference, 2*, 204–211.

C. Wallace, K. B. Korb, and H. Dai. (1996). Causal discovery via mml. *ICML, 96*, 516–524.

Chan, C. C. (2007). A framework for assessing usage of web-based e-learning System. *Innovative Computing, Information and Control*, 147–147.

F. Castro, A. Vellido, A. Nebot, and F. Mugica. (2007). Applying data mining techniques to e-learning problems. *Evolution of teaching and learning paradigms in intelligent environment*, 183–221.

H. Jeong and G. Biswas. (2008). Mining student behavior models in learning-by-teaching environments. *EDM*, 127–136.

J. Chen, Q. Li, L. Wang, and W. Jia. (2004). Automatically generating an e-textbook on the web. *Advances in Web-Based Learning–ICWL 2004*, 35-42.

J. Mostow and J. Beck. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. *Intelligent tutoring systems*, 353–362.

J. Tane, C. Schmitz, and G. Stumme. (2004). Semantic resource management for the web: an e-learning application. *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, 1–10.

L. P. Dringus and T. Ellis. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45*(1), 141–160.

M. Cocea, A. Hershkovitz, and R. S. Baker. (2009). *The impact of off-task and gaming behaviors on learning: immediate or aggregate?*

Muehlenbrock, M. (2005). Automatic action analysis in an interactive learning environment. *Proceedings of the 12 th International Conference on Artificial Intelligence in Education*, 73–80.

P. Reyes and P. Tchounikine. (2005). Mining learning groups' activities in forum-type tools. *Proceedings of th 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years!*, 509–513.

R. S. Baker and K. Yacef. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining, 1*(1), 3-17.

S. K. Mohamad and Z. Tasir. (2013). Educational data mining: A review. *Procedia-Social and Behavioral Sciences, 97*, 320–324.

Scott, J. (2012). *Social network analysis.* Sage.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

Ueno, M. (2004). Data mining and text mining technologies for collaborative learning in an ilms. *Proceedings. IEEE International Conference*, 1052–1053.

Universities. (2018). Retrieved June 19, 2018, from University Grants Commission - Sri Lanka: http://ugc.ac.lk/en/universities-and-institutes/universities.html

V. J. Hodge and J. Austin. (2014). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85–126.