MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 6, Issue 2, July-December 2021

Techniques of Educational Data Mining that Enable Evaluation

Farhana Karim
Professor
Department of CSE
Bangladesh University of Engineering and Technology
Corresponding Author's E-mail: *fhk.karim@gmail.com*

Abstract

Data mining has recently grown in favour among researchers. Data mining offers a variety of strategies and approaches for analysing data generated by various applications in various domains. Similarly, educational mining offers a method for examining educational data sets. Educational mining is concerned with establishing ways for uncovering information from educational data, and it aids in the extraction of hidden patterns and the discovery of new knowledge from vast educational datasets via the use of data mining techniques and tools. Educational mining knowledge may be utilised to make decisions at higher education institutions. This work is based on a survey of the literature on various data mining approaches, as well as specific algorithms such as classification, clustering, and so on. This research examines the efficacy of mining approaches with an educational data set for higher education institutions.

Keywords: *Data mining; Higher Education Data, Classification; Educational Mining; Clustering*

INTRODUCTION

"We live in the information era." We are, nevertheless, living in the data age. This rapidly increasing, publicly available, and massive quantity of data has truly ushered in the data era. Powerful and adaptable technologies are desperately needed to automatically extract meaningful information from massive volumes of data and convert such data into structured knowledge. Data mining arose as a result of this need. Data mining has made and

will continue to make significant advances in our transition from the data era to the approaching information age.

Data mining provides many methods and strategies for transforming raw data into valuable information by analysing massive data repositories from which new knowledge may be generated. Data mining, also known as Knowledge Discovery in Datasets (KDD), is the study of finding new and possibly usable information in huge databases. Data mining is used in a variety of fields, including marketing, healthcare, and law enforcement. Educational Mining (EM) is one of the application domains of Data Mining methods.

Data for education EDM's goal is to analyse such data and tackle educational research difficulties. Mining for Education

Educational mining is a new approach introduced by data mining. The latter's ideas are used to data linked to the subject of education in educational mining. It is the process of reorganising the information gathered by educational institutions. Educational mining is the process of analysing hidden data from educational settings using innovative technologies to gain a better knowledge of students and the context in which they learn.

Educational mining provides several tools and algorithms for analysing data trends. In EM, data is collected during the learning process and subsequently analysed using techniques from statistics, machine learning, and other data mining ideas. Various data mining approaches such as classification, clustering, rule mining, and so on have been addressed in order to extract hidden information from data derived from educational systems in order to generate better judgments in educational systems.

The academic's responsible and educators worked upon the educational system to strengthen the performance of students. In this diagram it is shown that educators want to design the educational system then plan to build that system and substantially maintain that educational system. Educational systems include traditional classrooms and some innovative learning methods like e-learning system, intelligent and adaptive web based educational system etc. The data set can be extracted from students as students are directly connected with educational system.

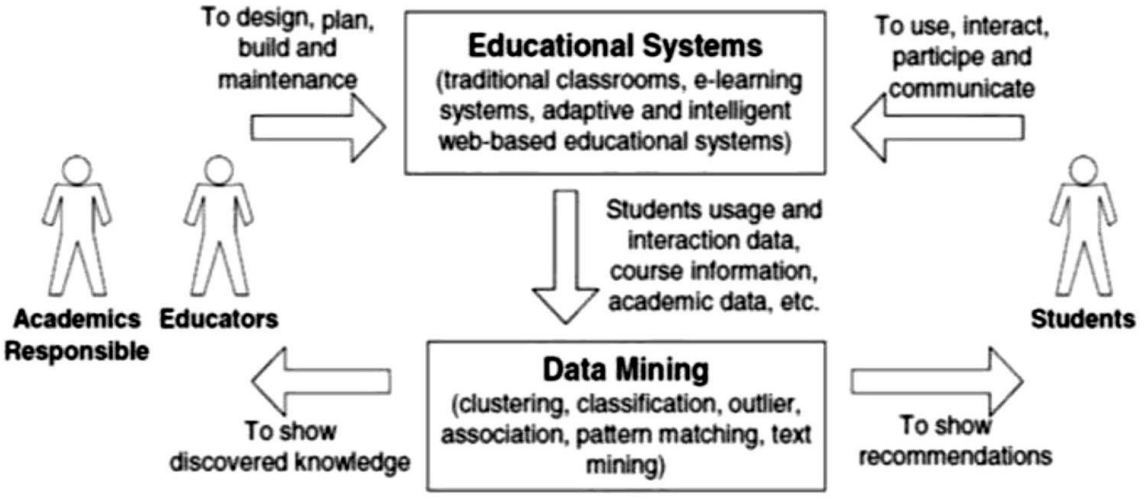

Figure 1: Educational Mining Process

Now the data is given as input to data mining methods and in result it produces guidance to students and to extract new knowledge to the educators by using various data mining techniques like clustering, classification, pattern matching etc.

Goals of Educational Mining

- Baker and Yacef describe the following four goals of Educational Mining:
- Predicting student's future learning behaviour
- Discovering or improving domain models
- Studying the effects of educational support
- Advancing scientific knowledge about learning and learners

1. Predicting student's future learning behaviour - With the use of hidden knowledge in student modelling, this goal can be achieved by creating student models that incorporate the learner's characteristics, including detailed information such as their knowledge, behaviour's and motivation to learn.

2. Determining or improving domain models - Through the several methods and applications of EM, discovery of new models and explore them to make appropriate improvements to existing models is possible.

3. Studying the effects of educational support - It can be achieved through learning systems.

4. Advancing practical knowledge about learning and learners - By constructing and incorporating student models, the field of EM research and the technology and software used.

Phases of Educational Data Mining

Educational Mining is concerned with transformation of new extracted knowledge from the raw data collected from educational systems. The data relevant to educational systems is collected which is to be mined from different educational system resources i.e. from course management system (different institutes), E-learning environment, web based data (i.e. YouTube, twitter)which is relevant to students activities during learning process(i.e. their academic grades, students posts on social networking sites etc).

EM generally consists of four phases:

1. The first phase of educational data mining is to find the relationships between data of educational environment. The purpose of constituting these relationships is to utilize these relationships in various data mining techniques like classification, clustering, regression etc.
2. The second phase of educational mining is validation of discovered relationships between data so that over fitting can be avoided.
3. The third phase is to make predictions for future on the basis of validated relationships in learning environment.
4. The fourth phase is supporting decision making process with the help of predictions.

RELATED WORK

This section concisely explains the literature review of various surveys and their comparative studies on the classification algorithms applied on educational mining.

In year 2010 Z. J. Kovacic explained a case study on educational data mining to identify upto what extent the enrolment data can be used to predict student's success. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

In year april 2011 Bhardwaj and Pal conducted the study on the student performance and found that the factors like students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance.

In year 2011 Pandey and Pal accomplished study on the student performance established by selecting 600 students from different colleges. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

In year 2012 Surjeet Kumar Yadav et al. have explained Decision tree algorithms on students' data set and found C4.5 can learn effective predictive models from the student data and gives the better accuracy of classification.

In year april 2012 Sonali Agarwal et al. describes the implementation by applying classification algorithms on educational data and found SVM classifier LIBSVM with Radial Basis Kernel has been taken as a best choice for data classification in her studies.

In year December 2012 Surjeet Kumar and Brijesh Bharadwaj have done comparative analysis on the decision tree classification algorithms and found CART algorithm is classifying the First, Second, Third class and Fail students with high accuracy.

In april 2013 Trilok Chand Sharma et al. discussed about the classification algorithms and explained how to run those classifiers for the selected dataset and found decision tree gives better performance and high accuracy.

In year 2013 V.Ramesh has done an analysis on classification algorithms for redicting student's performance and found Multi-Layer Perceptron predicted the performance better than the others and also found that parents' designation plays a vital role for predicting their grades.

In year July 2013 Sumit Garg et al. analysed various classification algorithms on educational data set and explained thorough details of the implemented algorithms and suggested some good algorithms to work on students' data set to predict their future performance.

In year 2014 Abdul Hamid and Amin analysed and compared students' enrolment approval using classification algorithms and found C4.5 gives high accuracy and lowest absolute errors.

In year 2015 Amirah mohamed Shahiri et al., have done a review on predicting students performance in data mining techniques and found classification algorithms predicts the performance better than other techniques in data mining and

C4.5 is highly used to by the researchers for predicting student's performance.

In year January 2015 Pooja Thakar et al., broadly analysed many papers on educational mining which compared the data mining technique predicts student's performance and found the attributes which are highly correlated with the student's performance.

In year 2015 C. Anuradha and T. Velmurugan selected classification algorithms and tested on students' data set and found that Classification of the students based on the attributes reveals that prediction rates are not uniform among the classification algorithms and also show classification algorithms works differently depends on the selection of attributes.

In year 2015 Sagar Nikam has done the comparative study of classification algorithms. Analysis of classification algorithm says each algorithm has its own merits and demerits and the techniques have to be selected based on the situation.

In year June 2016 R. Sumitha, and Vinothkumar [14], analysed and compared Classification algorithms on students' data set and found J48 gives better accuracy of 97%.

EDUCATIONAL MINING METHODS

EM not utilizes only data mining techniques like classification, clustering, and association analysis, but also applies methods and techniques drawn from the variety of areas related to EM (statistics, machine learning, text mining, web log analysis, etc.).There are various methods of educational mining but all kind of methods lie in one of following categories:

1. **Prediction:** The objective is to build a model which can interpret a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Predictions methods are classification, regression (when the predicted variable is a continuous value), or density estimation (when the predicted value is a probability density function).

2. **Classification:** It classifies a data item into some of several predefined categorical classes .The algorithm used for classification are:
   - Decision tree
   - Naive biased classification
   - Generalized Linear Models (GLM)
   - Support vector machine etc.

3. **Clustering:** In clustering technique, the data set is divided in various groups, known as clusters. As per clustering phenomenon, the data point of one cluster and should be more similar to other data points of same cluster and more dissimilar to data points of another cluster. There are two ways of inception of clustering algorithm: Firstly, start the clustering algorithm with no prior assumption and second is to start clustering algorithm with a prior premise.

4. **Relationship mining:** It is used for discovering relationships between variables in a dataset and encoding them as rules for later use. There are several types of association in mining techniques is available here which shows a better relationship between the attributes such as association rule mining (any relationships between variables), sequential pattern mining (temporal associations between variables), correlation mining (linear correlations between variables), and causal data mining (causal relationships between variables). In EM, basically relationship mining is an important concept to analyse the relationships between the student's on-line activities and the final marks and to model learner's problem solving activity sequences.

5. **Discovery with Models:** Its aim is to use a validated model of an event (using prediction, clustering, or knowledge engineering) as a component in further analysis such as prediction or relationship mining. It is used for example to identify the relationships between the student's behaviour and characteristics.

CLASSIFICATION ALGORITHMS USED IN EDUCATIONAL DATA

In this section, various classification algorithms used in educational mining is discussed here:

Classification Algorithms

Classification is one of the Data Mining techniques that are mainly used to analyse a given dataset. It is used to extract models that accurately define important data classes within the given dataset.

Classification is a two-step process.

Step 1: The model is created by applying classification algorithm on training data set

Step 2: The extracted model is tested against a predefined test dataset to measure the trained model performance and accuracy.

So classification is the process to assign class label from dataset whose class label is unknown.

Decision Tree

A decision tree is a decision support tool that uses a tree like graph and models and their possible consequences, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Curves from an internal node to its children are labelled with distinct outcomes of the test. Each leaf node has a class label associated with it. It is basically used for decision analysis and research operations etc.

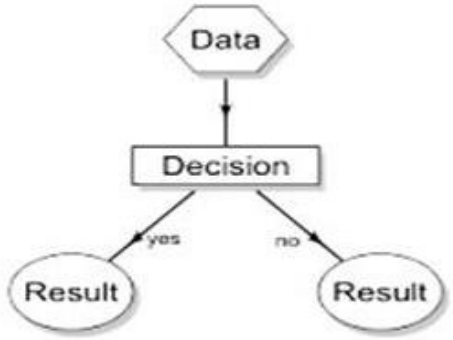

Figure 2: Decision Tree

A. C4.5

C4.5 is based on a supervised learning technique that builds a decision tree using both categorical and continuous characteristics. Essentially, it constructs decision trees from a given collection of training data is the same as ID3. To handle continuous attributes, C4.5 divides the attribute values from a particular training set into two smaller partitions based on the preset threshold, with all values over the threshold assigned to one child and the rest values assigned to another. It also deals with missing attribute values. C4.5 constructs a decision tree using Gain Ratio as an attribute selection metric. When there are multiple event values for an attribute, it minimises the bias of information acquisition. To begin, compute the gain ratio for each property. The root node will be the property with the highest gain ratio. It handles training data with various qualities and attributes with varying prices.

Forest Randomisation

A random forest is a collection of decision trees with some aspect of random choosing. Random forest works by constructing a number of trees to analyse the data, then combining all of the output from the trees and finally voting to achieve the final conclusion (search for classes that have the majority). The random forest provides excellent resilience for huge data sets, although it is more expensive than alternative approaches. Bayesian naiveté

The Naive Bayes approach is a statistical classification technique that is particularly useful when the input dimensionality is large. It is founded on the Bayesian theorem. It aids in predicting class membership probabilities, such as the likelihood that a given tuple belongs to a specific class. Based on the input, the Bayesian Classifier may compute the best potential output. It is also feasible to add fresh raw data during runtime and still have a useful probabilistic classifier. The influence of an attribute value on a given class is assumed to be independent of the values of the other attributes by naive Bayesian classifiers. This is known as class conditional independence. It is designed to simplify the computations required and is hence termed "naive".

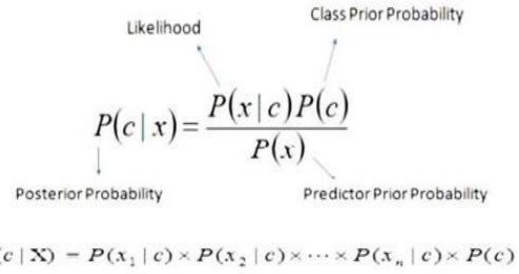

Figure 3: The Naive Bayes algorithm

- P (c|x) is the posterior probability of a class (target) given a predictor (attribute) of the class.
- P(c) is referred to as the prior probability of class.
- P(x|c) is the probability of a predictor of a particular class.
- P(x) is the class predictor's prior probability.
- Class (c) is another word for independent of the values of the other predictors.
- Perceptron with Multiple Layers

Multi-Layer Perceptron

This classification approach is a multi-layer perceptron feed-forward artificial neural network model that translates sets of input data onto a set of acceptable output data. As the name implies, it is made up of numerous layers of nodes in a directed network, with each layer completely linked to the one before it. This type of networks' schematic layout includes, in addition to the input and output layers, one or more extra levels known as hidden layers. The middle layer, also known as the hidden layer, performs intermediate processing before passing the input to the output layer.

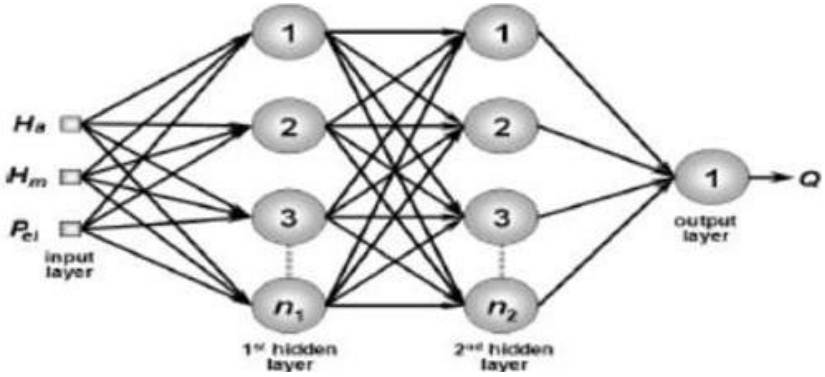

Figure 4 depicts a multi-layer perceptron.

Support Vector Machine (SVM)

Support Vector Machines are a type of classification technology that is based on the notion of decision planes, which determine decision boundaries. To translate the original training data into a higher dimension, this classifier approach employs nonlinear mapping. It seeks the linear optimum dividing hyperplane (i.e., a "decision boundary" separating tuples of one class from those of another) within this new dimension. Using support vectors ("important" training tuples) and margins, the SVM determines the optimal separable hyperplane (defined by the support vectors). SVM can handle many continuous and categorical variables and can perform regression and classification tasks. The kernel types in the support vector machine operator include dot, radial, polynomial, neural, and so on.

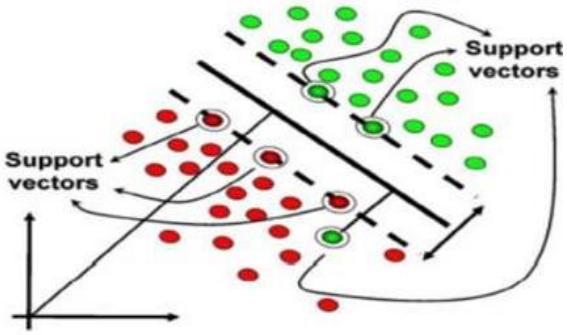

Figure 5 shows a Support Vector Machine

CONCLUSION

In this study, we highlighted how data mining techniques on educational data might be a beneficial strategic tool for higher education administration in tackling the critical problem of increasing the quality of educational processes. This study presents data mining techniques used in the educational area and provides a brief introduction of Educational mining, its aims, stages, and existing classification approaches.

Our research study will be useful to researchers who want to get insight knowledge in educational data mining in order to pick educational data mining strategies using their algorithms based on their research topic. As a result, we can conclude that this article will give a useful overview of present solutions for education data mining approaches, as well as their accuracy and benefits.

REFERENCES

1. Trilok Chand Sharma and Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 4, April 2013.
2. B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.
3. Surjeet Kumar and Brijesh Bharadwaj, "Data Mining Applications: A comparative Study for Predicting Student's performance", international journal of innovative technology & creative engineering, Vol.1 no.12, December 2012.
4. Pooja Thakar, Anil Mehta, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", International Journal of Computer Applications, Volume 110 – No. 15, January 2015.
5. C.Anuradha and T. Velmurugan, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance", Indian Journal of Science and Technology, Vol 8(15), 2015.
6. Amirah mohamed Shahiri, Wahidah Husain, "A Review on Predicting Student's Performance Using Data Mining Techniques", ELSEVIER, Volume 72, Pages 414-422, 2015.
7. Sagar Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms", Oriental journal of computer science &technology, Vol. 8, Pgs. 13-19, 2015.
8. Abdul Hamid and Amin, "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining",

Workshop on Interaction Design in Educational Environments, ISBN: 978-1-4503-3034-3, 2014

9. Sumit Garg and Arvind K. Sharma "Comparative Analysis of Data Mining Techniques on Educational Dataset", International Journal of Computer Applications (0975 – 8887), Volume 74– No.5, July 2013.
10. Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science & IT Education Conference, 2010
11. R. Sumitha and Vinothkumar, "Prediction of Students Outcome Using Data Mining Techniques", International Journal of Scientific Engineering and Applied Science, Volume-2, Issue-6, June 2016
12. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufman, 2006.
13. C. Pornnapadol, Children who have learning disabilities, Child and Adolescent Psychiatric Bulletin Club of Thailand, October-December, 2004, pp.47-48.
14. V.Ramesh, "Predicting Student Performance: A Statistical and Data Mining Approach", International Journal of Computer Applications, Volume 63– No.8, February 2013.
15. Sonali Agarwal and Pandey, "Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e- Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.
16. Surjeet Kumar Yadav and Saurabh Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 2, 2012.
17. U. K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", International Journal of Computer Science and Information Technology, Vol. 2-No.2, 2011.
18. M.J. Berry and G.S. Linoff, Data Mining Techniques: For Marketing, Sales and Customer Relationship Management, Wiley Publishing, 2004.
19. U.M. Fayyad and K.B. Irani, Multi-interval discretization of continuous valued attributes for classification learning, Proc. 13th

Int. Joint Conf. on Artificial Intelligence, 1993, pp.1022- 1027.

20. W. Jirapanthong, Classification model for selecting undergraduate programs, Proc. 8th Int. Symp. On Natural Language Processing, 2009, pp.89-95.
21. S. Wongpun and A. Srivihok, Comparison of attribute selection techniques and algorithms in classifying bad behaviors of vocational education students, Proc. IEEE Int. Conf. On Digital Ecosystems and Technologies, 2008, pp.526-531.
22. K. Waiyamai, C. Songsiri, and T. Rakthammanon, The use of data mining techniques to improve quality of engineering graduates, NECTEC Technical Journal, Vol.3, No.11, 2001, pp.134-142.
23. N. Sivaram and K. Ramar, Applicability of clustering and classification algorithms for recruitment data mining, Int. Journal of Computer Applications, Vol.4, No.5, 2010, pp.23 28.
24. A.L. Blum and P. Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence, 1997, pp.245- 271.
25. X.B. Li, J.R. Sweigart, J.T.C Teng, J.M. Donohue, L.A. Thombs, and S.M. Wang, Multivariate decision trees using linear discriminates and tabu search, IEEE Transactions on Systems, Man and Cybernetics, Part A, Vol. 33, No. 2, 2003, pp.194–205.
26. R. Sumitha and Vinothkumar, “Prediction of Students Outcome Using Data Mining Techniques”, International Journal of Scientific Engineering and Applied Science, Volume-2, Issue-6, June 2016.
27. Baker, R.S.; Yacef, K (2009). "The state of educational data mining in 2009: A review and future visions". JEDM-Journal of Educational Data Mining 1 (1): 2017.