Impact of Big Data Technologies in Education Management

Shamik Palit
Manipal Academy of Higher Education Dubai, United Arab Emirates

Chandrima Sinha Roy
Eminent College of Management and Technology Kolkata, India

Abstract: Big Data Technology (BDT) and Analytics have gained immense recognition in recent years. BDT plays an essential role in various sectors. This study intends to provide a review of BDT in the education sector which includes analyzing, predicting learner's results based on behavior patterns, assessing their performance regularly. Education institutions are beginning to use analytics and techniques for improving the services they provide and to enhance learner's performance and retention. BDT in education involves Data Mining, Data Analytics. This study also aims at investigating the techniques referred to as Educational Data Mining (EDM) and Learning Analytics (LA) influencing online learning systems. Thus, this research study will examine the field of EDM and LA which give an effective understanding on student's learning methods and identify their educational outcomes.

Keywords: Education management, Big data, Higher education

Introduction

Big data has become a buzz word in recent years as education, entertainment, communication is occurring over the web, thus generating a humungous amount of data. Individual in different sectors contributes to generating of big data. Data can be generated from heterogeneous data sources such as social media, email, transactions, etc. in the form of text, image, audio, video or a possible combination of these forms. To handle heterogeneity aspects of big data the traditional data mining techniques need to be upgraded. Commercial entities have led the path of developing techniques to gather insights from the massive data generated to identify likely consumers of their products, in refining their products to better fit consumer needs. More recently, researchers and developers of online learning systems have begun to explore analogous techniques for gaining insights from learner's activities online. The advancement of BDT has facilitated education with various types of teaching, learning and assessment methods that can be achieved in classrooms or virtualized environments. The learners can receive instant feedback on the content they are learning based on big data analytics. BDT can analyze the overall performance of a class at a macroscopic level and can analyze each student's or learner's performance to find strengths and weaknesses. Then accordingly educators can take decisions on weak points of students to enhance

Palit, S. & Roy, C. S. (2021). Impact of big data technologies in education management. In M. Shelley, I. Chiang, & O. T. Ozturk (Eds.), *Proceedings of ICRES 2021-- International Conference on Research in Education and Science* (pp. 194-207), Antalya, TURKEY. ISTES Organization.

their performance.

More recently, researchers and developers of online learning systems have begun to explore analogous techniques for gaining insights from learner's activities online. The advancement of BDT has facilitated education with various types of teaching, learning and assessment methods that can be achieved in classrooms or virtualized environments. The learners can receive instant feedback on the content they are learning based on big data analytics. BDT can analyze the overall performance of a class at a macroscopic level and can analyze each student's or learner's performance to find strengths and weaknesses. Then accordingly educators can take decisions on weak points of students to enhance their performance.

Analysis of large educational datasets can be done by mainly two techniques: Educational data mining (EDM) and Learning Analytics (LA). These techniques respond to the event-based analysis related to education policies and practice. EDM is a DM technique applied on educational data sets. It aims to better understand students in terms of their learning pattern. EDM applies a combination of techniques such as machine learning, data mining to understand the research, learning issues in the educational sector (see Figure 1).

LA is the collection, analysis and reporting of data about learners. LA analyzes the large datasets and provides feedbacks that have an impact on students, instructors, and the learning process. There are quite a few differences between EDM and LA; Researchers in EDM rely more on classification and clustering, whereas in LA researchers use statistics, visualization, Social Network Analysis, sentiment analysis, influence analysis.

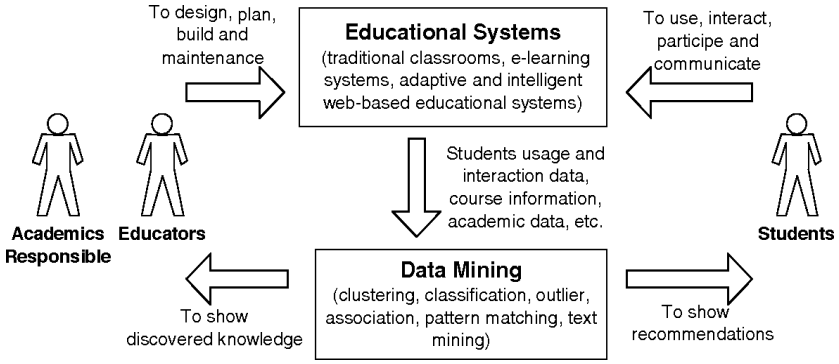


Figure 1. Application of Data Mining in Educational Systems

Literature review that has been done author used in the chapter "Introduction" to explain the difference of the manuscript with other papers, that it is innovative, it are used in the chapter "Research Method" to describe the step of research and used in the chapter "Results and Discussion" to support the analysis of the results [2]. If the manuscript was written really have high originality, which proposed a new method or algorithm, the additional chapter after the "Introduction" chapter and before the "Research Method" chapter can be added to explain briefly the theory and/or the proposed method/algorithm [4]. The use of e-learning systems has grown exponentially in the recent years, by the fact that neither the students nor the teachers are bound to any specific location and that is the form of computer-based education is virtually independent if hardware platform. LMS

are becoming much common in universities, community colleges, schools and are even used by instructors/educators to add web technology to their courses. LMS produce content material, prepare assignments and tests, engage in discussions, manage distance classes etc. LMS have a collection of vast amounts of information which is very valuable for analyzing student's behavior, very valuable source of educational data. They provide a database that stores personal information about the users, academic results of the students and their interaction data. Due to vast quantities of data generated on a daily basis, managing manually becomes difficult. These platforms do not have specific tools to allow education track and assess all student/learner's activities while they evaluate the structure and contents of the course. The use of data mining is very efficient and necessary to help educators, courseware authors to improve and enhance the educational systems through EDM &LA.

Problem Definition and Objective

The aim of student in an educational institute is to focus on academics. Some students might be good academics while some may perform poor. There can be several reasons on poor performance of the student such as difficulty in understanding the course, neglecting the course, their friend circle or may be some other several reasons. To understand this issue and track the students on regular basis can help in resolving issue up to certain extent, which can be done through EDM and LA. To resolve the issue of students performing poor, we intend to study the field of EDM &LA:

- An understanding of Educational Data Mining (EDM) applied to large data sets of students generated in the educational sector.
- An understanding of Learning analytics (LA) and how it is applied in the education sector.
- The benefits of LA and EDM and what factors have enabled these approaches to be adopted.

Background

Students from various locations using online platforms for learning purposes, generate vast quantities of data on a daily basis, it is difficult to manage manually or with traditional techniques. Therefore, it gets harder for an instructor to extract useful information as there are large number of students each generating large amounts of on a regular basis. The data mining techniques have been applying on large volumes of educational data sets to extract useful information required by the instructors which accordingly instructors can monitor the student performance, these data mining techniques applying on educational data sets are known as EDM. "EDM is an emerging interdisciplinary research area that deals with application of DM techniques on educational data.

Classification is most frequently studied by DM and ML researchers. It consists of predicting the value of a categorical attribute based on the values of other attributes that is the predicting attributes. In classification, it is an approach of supervised learning. Classifier from set of correctly classified instances known as the training set. This classifier is used in algorithms directly. The other set known as testing set is used to measure the quality of the obtained classifier after the learning process. Different types of models can be used to represent

classifiers obtained from the training set. One of the algorithm types which can be used is decision trees:

Decision trees: It's a hierarchical structure where set of conditions are organized that contains zero or more internal nodes and one or more leaf nodes. Arcs with labeled node to its children are labeled with different outcomes of the test at the internal node. The decision tree is a predictive model in which an instance is classified by following a path of conditions which satisfy from the root node of the tree reaching the leaf node, which corresponds to the class label. Different and well-known classification algorithms are ID3, C4.5, CART etc (see Figure 2).

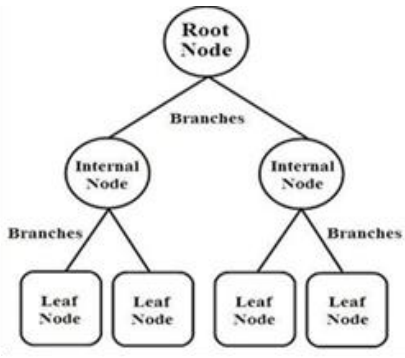


Figure 2. Decision Tree

Clustering is a DM technique which identifies data points that are similar in some respect so that a full dataset can be split into various categories of small datasets. Unlike classification here the training data set is not provided, hence it is an unsupervised learning. Some of the well-known clustering algorithms are K-Means, hierarchical clustering etc (see Figure 3).

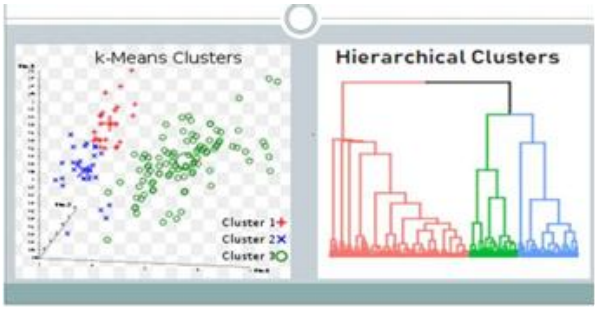


Figure 3. Clusters

Literature Review

The Knowledge gathered from research papers were on various data mining algorithms used on educational data sets and how they process educational data sets. The following are summary of citations of the algorithms that would detail in the further chapters of the report.

ID3: In,[6], Kalpesh Adhatrao &Aditya Gaykar define the process ID3 Algorithm and how it is applied on educational data sets. It describes the parameters which decide the root nodes, internal nodes and leaf nodes of

the decision tree.

C 4.5: In [7], T.Miranda Lakshmi, A.Martin define the process of C4.5 Algorithm and how it applies on educational data sets. In the algorithm it specifies the parameters that choose the nodes of the decision tree.

K-Means: In [8], Velmurugan T, C Anuradha define the process of K-means Clustering Algorithm and how it impacts in the educational sector, how clusters are formed and grouping based on the clusters is done.

Hierarchical Clustering: In [8], Velmurugan T, C Anuradha define the process of Hierarchical algorithm can process the data sets in the education sector, the steps and formation of cluster is being described by them.

Methodologies

In this study, methods used to carry out the reseach study are classification and clustering algorithms and one comparative analysis between the classification algorithms and other between clustering algorithms. The comparative analysis will be done based on: The classification technique involves learning and classification if data. Most of the frequently used DM method, which develops the cases and assign data set to the classes. The target data is evaluated by classification algorithm. In classification method, the test data are utilized to evaluate the efficiency of the classification rules. If the rules are acceptable, it can be utilized to new data sets.

Decision Tree is used to predicting the student's academic performance. Example: Classifying learners according to their interactions with course content (video lectures and assessment) in learning activities and forecast student performance based on their interacting behavior.
The two algorithms of classification to be discussed are:

- ID3-This is a decision tree algorithm introduced in 1986 by Quinlan Ross. It is based on Hunts algorithm
- C4.5-This decision tree algorithm is a successor of ID3 and based on Hunts algorithm.

The Clustering is an iterative process of discovering knowledge. The techniques find classes an assign the object to a desire class. Example: In terms of educational sector grouping students based on their learning and interaction patterns and grouping users for purposes of recommending actions and resources to similar users
The two algorithms of clustering technique to be discussed are:

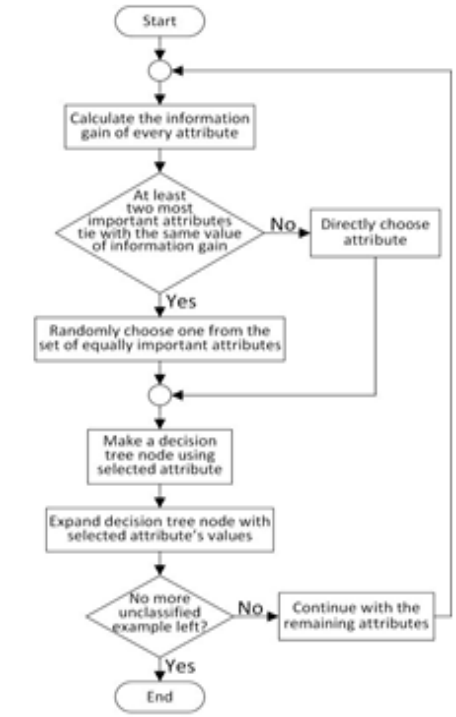
- K-Means
- Hierarchal Clustering

DM Algorithms

- ID3 uses information gain measure to choose splitting attribute.

- It accepts categorical values in building a tree model and doesn't give accurate result when there is noise, hence to remove the noise preprocessing technique is required.
- To build the decision tree, information gain is calculated for each and every attribute and the attribute with the highest information gain is selected to designate as the root node.
- The attribute is labelled as a root node and the possible values of the attribute are represented as arcs.
- All the possible outcome instances are tested to check whether they are falling under the dame class or not.
- If all the instances are falling under the same class, the node is represented with single class name otherwise choose the splitting attribute to classify the instances.

1. ID3



Flowchart of the traditional ID3 algorithm

Formula of Entropy

Given probabilities $p_1, p_2, \ldots, p_s$, where $\sum p_i = 1$, Entropy is defined as

$$H(p_1, p_2, \ldots, p_s) = \sum - (p_i \log p_i)$$

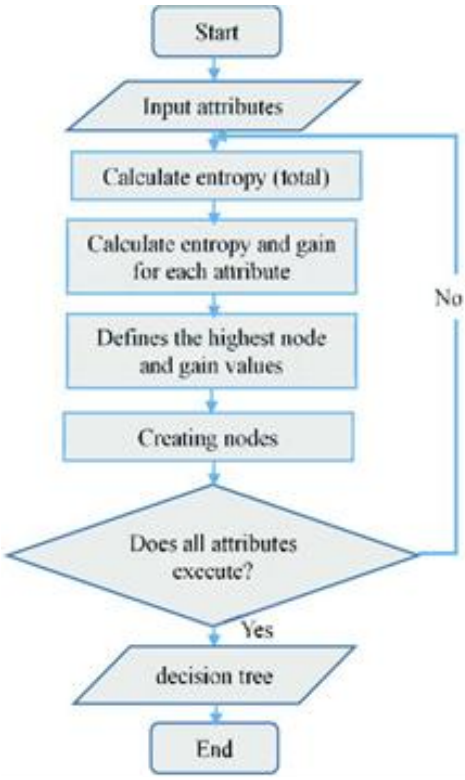
Entropy finds the amount of order in a given database state. A value of H = 0 identifies a perfectly classified set. In other words, the higher the entropy, the higher the potential to improve the classification process.

Formula of Information Gain:

ID3 chooses the splitting attribute with the highest gain in information, where gain is defined as difference between how much information is needed after the split. This is calculated by determining the differences between the entropies of the original dataset and the weighted sum of the entropies from each of the subdivided datasets. The formula used for this purpose is:

$$G(D, S) = H(D) - \sum P(D_i)H(D_i)$$

2. C 4.5



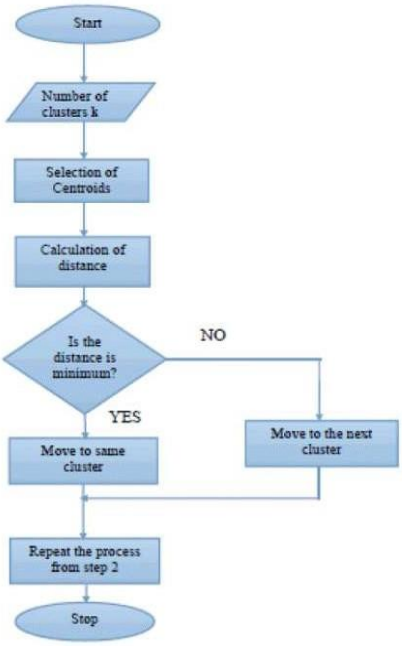
C4.5 handles both categorical and continuous values

To handle the continuous attributes, C4.5 splits the attribute value into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child.
Handles missing attribute values.

It uses gain ratio as an attribute selection to measure to build a decision tree. Firstly, calculate the gain ratio of each attribute The root node will be the one whose gain ratio is maximum.

This algorithm uses pessimistic pruning to remove unnecessary branches in decision tree to improve the accuracy of classification.

3. K-Means Clustering



In this algorithm K data elements are selected as initial centers and Euclidean distance formula is used to calculate the distance between the selected centroid and other data elements and then same procedure is followed iteratively.

Firstly, select the no of 'c' cluster centers. (Here fixed no of clusters is used).

Initial cluster center is determined for each of the c clusters either by software or the researcher.

Then distance between each data point and cluster center is calculated using Euclidean distance formula.

Assign the data point to that cluster center whose distance from the cluster center is minimum as compared to all cluster centers

Recalculate the new cluster center using the formula:

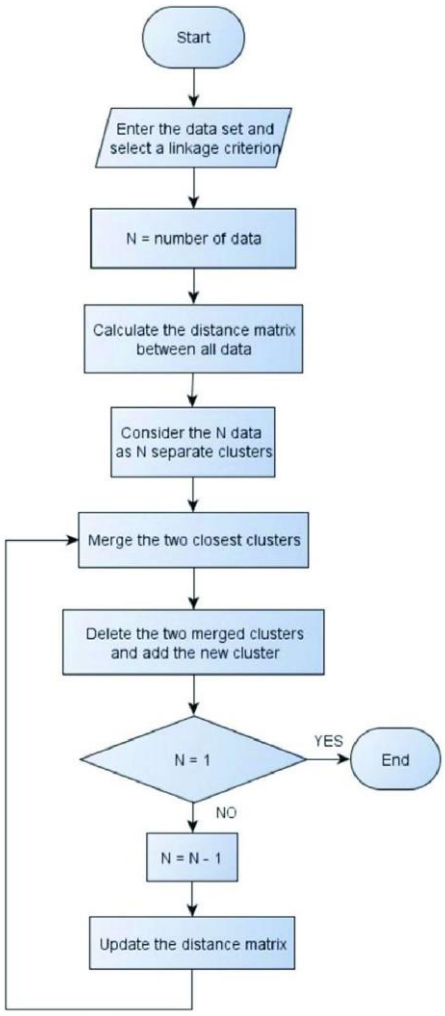
$$vi = (1/ci)\sum_{j=1}^{ci} xi$$

Where 'ci' represents the number of data points in i^{th} cluster.

Recalculate the distances of each data point with new cluster centers.

If there is no reassigning of the data points then stop, otherwise repeat the calculate the Euclidean distance step.

4. Hierachchial clustering



Hierarchical Clustering is basically set of nested clusters organized in a hierarchical tree.

Assign a cluster to each item, such that N clusters for N items.

Find and merge the pair of clusters which are closet to each other.

Calculate the distances between new and each of old clusters.

A) Start with the disjoint clustering 1(0) =0 and sequence number n=0.

B) In the current clustering, now find the least dissimilar pair of clusters say pair (A), (B), according to d[(A),

Increment the sequence by n=n+1 and merge clusters A sand B into single cluster to form the next clustering n.

Set the level of this clustering 1(n)=d[(a), (b)]

The next step is to update the proximity matrix, M, by deleting rows and columns corresponding to clusters A and B and adding a new row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted by (A, B) and old cluster k is defined in this way[(k), (a, b)] =min d[(k), (a)], d[(k), (b)].

If all the objects are in one cluster then stop the process, else go to step to find and merge the pair closer to each other.

Implementation Details

Moodle is one of the most widely used online learning environments by educational institutions. Educators inexperienced in DM can carry out basic Dm analyses on log records they obtained via Moodle LMS. This way educators will be able to obtain data—driven information on the status of the learning environment and students, while researchers will be able to seek answers for research questions regarding online learning. According to available data, it is used by over 138 million registered users in 230 countries. Furthermore, it is distributed for free with an open source code. It offers educators effective tools for providing course materials to students and organizing online learning activities. Moodle does not store simple text files. It registers the logs, and all information in a relational database. It allows to get full reports on the activities of a unique student or of all students for a specific activity or course. However, all this data is usually raw, without any form of intelligent processing, is uses have sued different told for Moodle data analysis.

The EDM in Virtual Learning environment follows mainly four steps:

- Data collection: while the students use the system, information is collected and stored in the database. In Moodle, the data is collected in system logs.
- Preprocessing: after data collection, the data is transformed into suitable formats for analysis. Usually software is used for data preprocessing.
- Data mining: with the aim of developing a model and discovering useful patterns, the appropriate data mining algorithms are applied at this stage.
- Results evaluation: in this last step, educators interpret the obtained results and use discovered knowledge to improve the learning and decision-making process.

Moodle Predicta is a tool developed in Java that allows users to connect into any version of Moodle DB as well as different management systems.

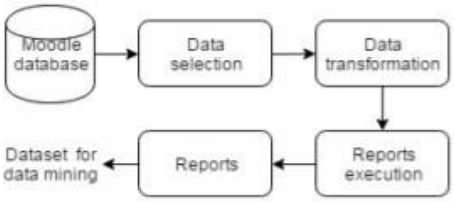
Moodle PREDICTA is divided into two parts:

1. **Visualization Module**

Allows users to have an overview of student behavior, interactions, personal data and academic performance. This module enables educators to evaluate the course structure and its effectiveness.

To bring the processed data to the user, this module is comprised of mainly four steps:

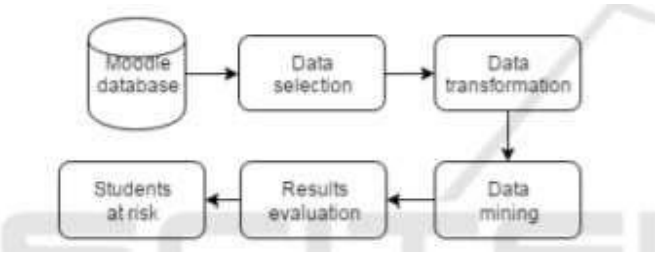
- Data Selection: Here the data (Moodle DB) is selected, according to user's requirements.
- Data Transformation: After selection, the data is gathered and transformed/discretized.
- Reports execution: In this stage, the reports are generated.
- Reports: The reports are presented to the user.



Once the user has selected the desired DB, he is asked for specific course on next screen. To make the selection Moodle Predicta creates a hierarchical structure of courses, according to categories and subcategories in selected DB, When the course is selected, a new screen is presented to the user. The selected attributes will compose of reports and quizzes, logs, assignments and grades. Once the user makes the selection, Moodle PREDICTA generates a file as the final result, which can be in mainly three formats: HTML, CSV, ARFF.

2. Prediction Module

The prediction module allows teachers and tutors to identify students not following classes, may abandon courses before the end, making it possible to take some preventive action.

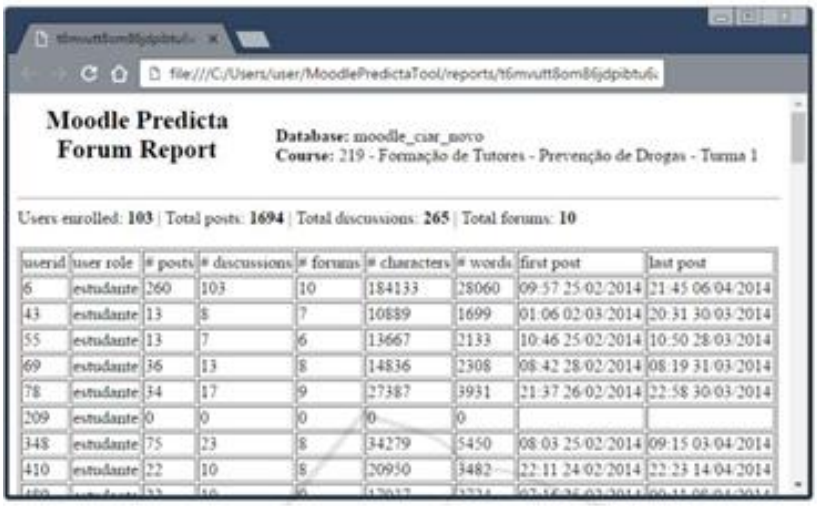


To undertake the students' prediction performance, this module is composed of mainly five steps:

- Data selection – in this step, the data is selected, from attributes data describe students' behavior, interactions
- and grades.
- Data transformation – the data is gathered and transformed/discretized for data mining.
- Data mining – in this stage, the data is used in decision tree models.
- Results evaluation – the results presented by data mining are interpreted and evaluated automatically.
- Students at risk – Students at risk are listed to user.

Once the course is selected, a, Moodle Predicta prepares the data, in a preprocessing phase (cleaning, integration, transformation and reduction), and generates an ARFF file. After connecting to the WEKA data mining API, the decision tree algorithms are executed with standard parameters for listing of students at risk of failing. Students whose behavior, interactions, and performance, is similar to those students from the training dataset that have failed will be defined as "at risk". Teachers can then follow up on these students to confirm

their situation and take some.



Comparative Study and Results Obtained

Comparative study between ID3 and c4.5 classifiers are:

	Splitting Criteria	Missing Values	Attribute type	Speed
ID3	Information gain	Doesn't handle it	Handles categorical values	Low
C4.5	Gain ratio	Can handle it	Handles both categorical and Numeric value	Faster than ID3

Comparative study between K-means and hierarchical clustering are:

Properties	K-Means	Hierarchical Clustering
Clustering Criteria	It is well suited to generating globular cluster	Use a distance matrix as Clustering Criteria
Category Data	K-means can be used in categorical data is first converted to numeric by assigning rank.	Applies categorical data and due to its complexity, a new approach for assigning the rank value to each categorical attribute.
Sensitive to noise	K-Means is sensitive to nose in dataset.	Comparatively less sensitive to noise in the dataset.
Execution time	Increases time of execution	Better performance than K-Means
Dataset	Good for larger data sets	Relatively good for smaller data sets.

Conclusion

With the advancement of information and communication technologies, new and major challenges being created mainly because of huge volumes of data about student's activities, academic results and user's interaction being stored. However, this data can be explored and analyzed by knowing DM techniques and algorithms. These facts are the basis of the recent area of research educational data mining, that consists of DM technologies applying to data collected from educational institutions with the aim to discover patterns and useful information. The data mining processes are difficult and need previous knowledge to be applied successfully. Moreover, the data needs to be correctly selected, prepared and the result of process requires evaluation and interpretation. In this study, the DM algorithms under classification and clustering techniques are studied to understand how they can apply on EDM. Also, in this study Moodle Predicta, an easy-to-use tool was presented. This software enables students follow up, selecting and preparing the Moodle data for two modules: (i) the visualization module, that generates reports for analysis purposes; and (ii) the prediction module, that integrated to WEKA data mining software, uses decision tree models to identify and list students at risk of dropout or failure.

References

[1] The main references are international journals and proceedings. All references should be to the most pertinent and up-to- Santosh Kumar Ray and Mohammed Sayed, "Applications of Educational Data Mining and Learning Analytics Tools in Handling Big Data in Higher Education: Trends, Issues, and Challenges", Khawarizmi International College, Al Ain, UAE, July 2018.

[2] Marviah Adib Hamiah, Sarfaraz N Brohi and Babak Bashari Rad, "Big Data Technology in Education: Advantages, Implementations and Challenges", Taylor University, Lakeside Campus, Selangor, Malaysia, July2018.

[3] Jiechao Cheng, "Data Mining Research in Education", Wuhan University, Wubei, China, March 2017.

[4] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method", Sadat Academy, Cairo, Egypt, 2015

[5] Velmurgun.T, "Clustering Algorithms in Educational Data Mining: A Review", D.G Vaishna College, January 2015

[6] Kalpesh Adhatrao, Aditya Gaykar," PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS" Navi Mumbai, Maharashtra, India Vol.3, No.5, September 2013

[7] T.Miranda Lakshmi, A.Martin, R.Mumtaj Begum, "An Analysis on Performance of Decision Tree Algorithms using Student Qualitative Data ",Bharathiyar University, Coimbatore, India, 201

[8] C. Anuradha, T. Velmurugan, "Clustering Algorithms in Educational Data Mining A review", Bharathiar University, Coimbatore, India, Vol 7. No.1 Pp.47-52, 2015

[9] Cristobal Romero, Pedro Espejo," Web Usage Mining for predicting final marks of student that use Moodle Courses", Cordoba University, Spain, May 2013

[10] Igor Moriera Felix, Ana Paula," Moodle Predicta: A Data Mining Tool for Student Follow Up", Instituto de Informatica, Universidad Federal de Goi ´ as, Goi ´ ania, Brazil, 2017

[11] Manju Kaushik, Bhawana Mathur, "Comparative study of K-means and Hierachical Clustering Techniques", JECRC university, Jaipur, India, Vol 2 Issue 6, June 2014