International Journal of Management, IT & Engineering
Vol. 8 Issue 6, June 2018,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Educational Data Mining: Predicting student's performance using clustering

Prof. Priya Chandran

Ms. Sanakhatun Shakirali Shaikh

Abstract

Educational Data Mining (EDM) is an emerging trend which is concerned about the effective use of huge data coming from educational systems to improve and optimize teaching-learning process, and also there are various techniques and methods provided to predict and analyze the student's behavior and their performance. In this research paper, we have developed a model to predict the student's performance by using clustering model with the help of k-means algorithm. We have focused on social media perspective data, which is strongest source of EDM and nowadays, it is more widely used to be as an important factor in the field of EDM.

Keywords:

Education Data Mining;
Educational Systems;
clustering model;
k-means algorithm;
Social media.

Author correspondence:

Prof. Priya Chandran
Email: priyaci2005@gmail.com
Ms. Sanakhatun Shakirali Shaikh,
Email: sana.shaikh.vashi@gmail.com

1. Introduction

In today's world, technology increasingly growing and the more technology enhancing, the concept of e-learning resources, educational software are exponentially growing and this produces the huge collection of datasets. This information is very helpful to discover the student's behavior and trends. It is not feasible to analyze the data collected from large repositories and data warehouses manually; it is effective for small databases but also becomes the bottleneck for large data. In this situation, data mining technique is used in education sector. Data mining provides the ability to analyze and predict data from multiple sources with different dimensions.

In Educational Data Mining (EDM) data is analyzed from various sources such as educational systems of institutions and schools, online courses, online assessments, virtual classrooms which generates huge quantities of very fine grained real time data on students learning that can be used for wide variety of things.

Data obtained from these sources are heterogeneous which can be further processed and analyzed for predicting the various activities of students which will helps to improve teaching learning process. Data for EDM are not restricted that it will obtained from only educational systems of institutes or schools but also from the social activities of the students. Students of this generation are mostly engaged with their social lives, even the students are more dependent on online resources to study which are available globally across

the world. There are many platforms from where students share their problems, ideas and innovations, contents, knowledge associated to studies. These data resides in educational databases are statistically analyzed [1] and give us information about students browsing data i.e. the websites which are frequently accessed by the students, resources that are downloaded from the web in order to study, the time spend by the student on webpage, interactions happened between the sites and students this will help to identify the student's behavior.

According to these scenarios mentioned above, reports are generated periodically (weekly, monthly, and annually).As a result of above analysis, various applications are available due to the predictions of the student's data available in huge databases or repositories. Always socio- economic status should have the influence over the student's academic achievements [2]. In this research paper we are proposing a clustering based approach to identify student's performance. We have collected students data based on personal, academic and social media attributes.

2. Research Method

Data mining is applied for very large datasets using the supervised and unsupervised algorithms [3]. EDM is the emerged area for research as the technology improved and is applied into education field. Different types of data originated from different sources is analyzed to identify behavior of the students and used for further research on various areas of education data mining [4].

EDM is an emerging application of data mining which concerned to generate specific type of data that is extracted from various educational environment to predict behavior and the pattern of student learning process and often to predict faculty's behavior [4], [5]. The technique called clustering is an important aspect in educational data mining which is used to group the students according to their behavior and performance [6]. Data generated in schools, colleges, universities, learning institutions providing traditional and modern forms and methods of teaching, as well as informal learning shall be analyzed with the help of EDM [4]. The above said issues regarding traditional education system are overcome by EDM in order to enhance students learning experience and profit is exponentially increased in business perspective [7]. C. Romero et al. [8] discussed application of DM in LMS. The authors described data mining process of e-learning data, step by step, as well as how to apply the main data mining techniques such as statistics, visualization, classification, clustering and association rule mining on MOODLE data.

Education is the future of our nation. So standard of the education should be advanced, which helps to increase the growth of the students and their career, to provide them better opportunities, and to explore their skills and make their future bright.

Data mining is the disciplinary field which covers multiple areas such as business, statistics, visualization, education, machine learning, pattern recognition, finance, artificial intelligence, etc. Earlier, retrieving, maintaining, prediction of patterns of students becomes a tedious work by using traditional techniques. Traditionally, educational data may manually obtained by the sources like interviews, focus group, questionnaires, attendance, tests, surveys, observations, and so forth. Using these techniques it is not an easy task to predict and analyze the student's behavior, teaching-learning process, and make the decision based on them. Availability of data is major issue. Data obtained by these techniques are not available anytime, anywhere across the globe so that data will not be optimized properly.

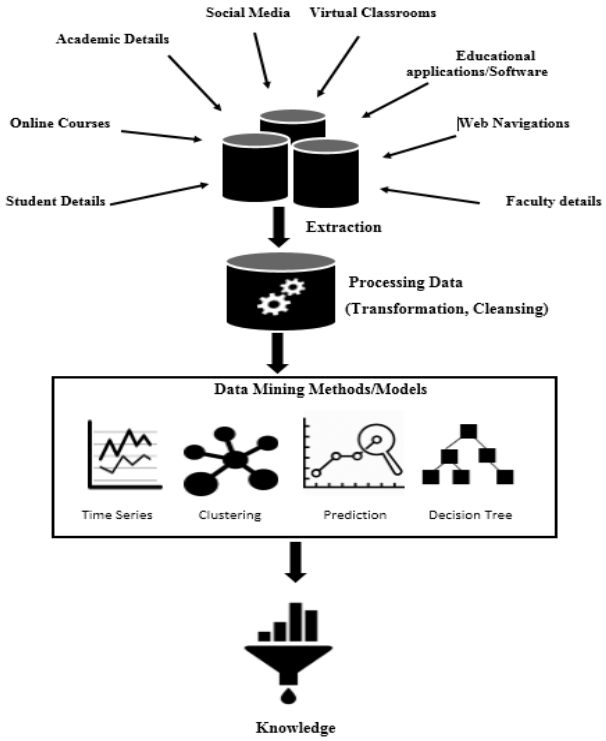

Figure 1. *Knowledge discovery with data mining models*

Figure 1 depicts the huge amount of different data is extracted from various sources (some of the sources are mentioned are e-learning, online course details, student's details, faculty details, academic details, session data, web navigation, virtual classrooms, educational applications, educational software's(such as Learning Management systems, Assessment Software's, Reference Software, and so on)[9], Web navigation, etc. Educational Data Mining provides techniques that allow us to standardize the collection of huge amount of data which exponentially growing across the globe real time dynamically. It also provides the facility to generate analytics of the data being collected from various sources and produce reports.

Since information technology widely covers the entire education sector, it is possible and easier to analyze the behavior and performance of students. In this section we briefly outline the Education Data Mining. Educational Data Mining is defined as the process of determining the pre-existing data resides within huge repositories, in order to obtain the knowledge which helps to predict student's behavior, their performance, and improvements in teaching and learning process. It is the study of learning the data from educational systems. Using EDM tools such as association rules, clustering, decision trees, time series analysis, prediction, and etc. the education management are able to plan and make decisions based on the activities going on currently to predict the future.

3. Results and Analysis

Clustering is defined as the grouping of unsupervised set of objects based on their similar characteristics [10]. The term cluster is defined as the group of objects with similar characteristics contained within clustering. Clustering is similar to classification, the only difference in it is that clusters are not predefined where in classification groups are predefined. Set of objects within a one cluster is similar to each other but different than set of objects resides within another clusters. Results obtained by clustering are dynamic due to Cluster model is broadly classified into following categories depicted by figure 2 [11].

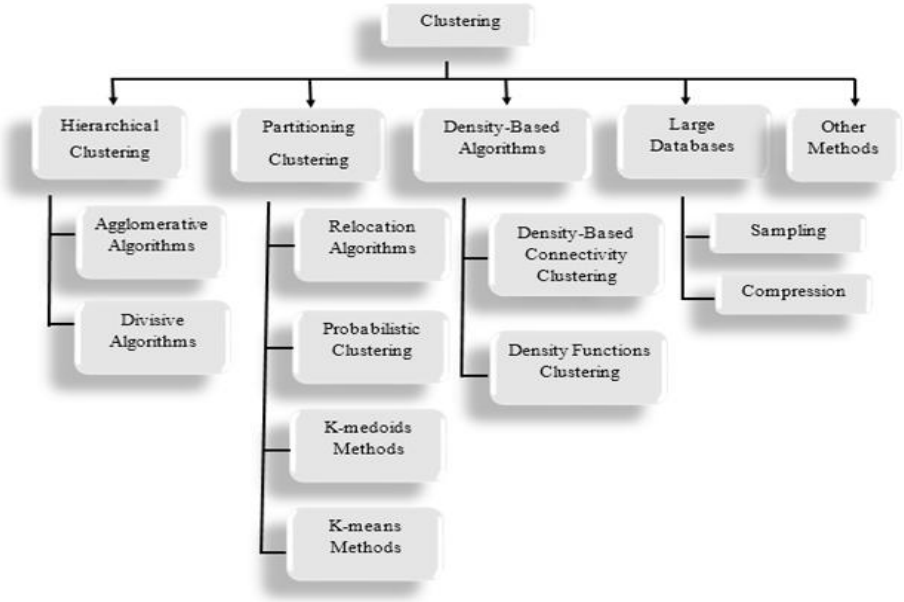

Figure 2. *Classification of Clustering in Data mining*

In this paper, we are using k-means clustering technique to identify student's performance. K-means methods are most frequently used clustering technique, which can partition the unlabeled [12] data sets into 'k' number of clusters automatically [9]. Here the term unlabeled data means that the available data set is currently does not belongs to any group i.e. it does not have preexisting groups available because as we are familiar that k-means is the type of clustering which contained unsupervised type of learning. K-means algorithm are simple to implement [13] due the reason that k-means provides the methods which provides approximate, scalable solutions. This algorithm is mainly used to discover the groups based on the number assigned to the k variable in order to represent the groups [12]. We are using Euclidean distance calculation technique of k-means clustering to discover the number of clusters in the data set. In this algorithm,

- Initial clusters are randomly chosen.
- Iteratively, items are moved among sets of clusters until the desired set is reached.
- High degree of similarity among elements in a cluster is obtained.

Analysis of data is most important part of decision making. In EDM, various data mining techniques and methods are used to predict and uncover hidden trends and patterns and making accuracy based predictions through a higher level of analytical sophistication in the process of counselling students.
To keep track of student's details, various types of educational software's are available which may varies in subjects [14] and also now for the ease of use and availability, IT companies associated to educational domain are building the educational applications that are beneficial for both students and teachers used as teaching-learning tools. These teaching-learning tools provide functionalities to both teacher and students. Due to these tools teachers are able to posts assignments, share lecture notes and presentation slides, tutorials, videos, and so on. In the recent scenarios, the educational companies integrate the education applications with the games. These games are the instructional games.

Most popular example of EDM is Moodle which is open source online learning management system and anything produced by them is freely available all over the globe. In Moodle faculties are able to give the assignments to the students, they can chat, organize aptitude tests, and much features are available that improves the learning-teaching procedure.

As we know, EDM is most widely used to evaluate and interpret teaching-learning process, student's performance and so on. Some of the activities are mentioned below to better understand about EDM. Predicting and analyzing useful data consists of performance, learning, interest, needs of students from raw data is very important in educational community. This extracted data involves various techniques available in data mining. In this paper we have taken used dataset which was collected from various students pursuing MCA.

We have considered some possible sources of Education Data Mining mentioned in the give table 1 below, through which, with the help of data mining tool known as WEKA. We have calculated the student's performance using clustering technique. We have used WEKA tool for analyzing data.

Table 1. Some possible sources of EDM to discover student's performance

Attribute	Description
Students personal details	
Age	Students age
Gender	Students gender
Status	Students marital status
Academic Details	
Year of enrollment	Students year of admission
Attendance	Attendance of the semester
Assignments	Assignment of each semesters based on subject.
Internal tests	Scores of internal test of each semester.
Lab	Practical lab scores
SGPA	Semester wise scores
Year of passing	Students year of passing
Internship	Students who got only internship
Placement	Final Placement
Events	Annual events
Social Media	
Educational related apps	Applications used by students for studies.
E-resources	Resources available for the students for studies.
Blogs	Interactions of the students and getting the knowledge
Forums	Ideas, innovations shared by students
Sharing of media	Photos, videos, pdf, and other media shared by students.

We grouped the students according to their semester's grades and their performance index should be depicted in the table 2. The table 2 represents the letter grades and their equivalent grade point applicable for MCA. Sample data set is shown in table below:

Table 2. Letter grades and their equivalent grade point applicable for MCA

Percentage of Marks Obtained	Letter Grade	Grade Point	Performance
80.00 and above	O	10	Outstanding
75.00-79.99	A	9	Excellent
70.00-74.99	B	8	Very Good
60.00-69.99	C	7	Good
55.00-59.99	D	6	Fair
50.00-54.99	E	5	Average
45.00-49.00	P	4	Pass
Less than 45.00	F	0	Fail

5. Result and Discussion

We applied the model on the academic result of the MCA students on semester basis. Result analyzed from the data set is described in the table 3 given below. We took three clusters, k=3 ,with cluster size of 50. The size of the first cluster is 12 and size of second and third clusters are 26 and 12 respectively. After two iterations of the data set we came to conclusion that the overall performance of the cluster size 12 is 7.38 while the overall performance of the second cluster having cluster size 26 is 4.98 and also the third cluster with cluster size 12 is 9.41. Result was predicted by implementing the k-means model. Figure 3 (a)- (d) shows the results of the study.

Table 3. K=3

Cluster	Cluster Size	Overall Performance
1	12	7.38
2	26	4.98
3	12	9.41

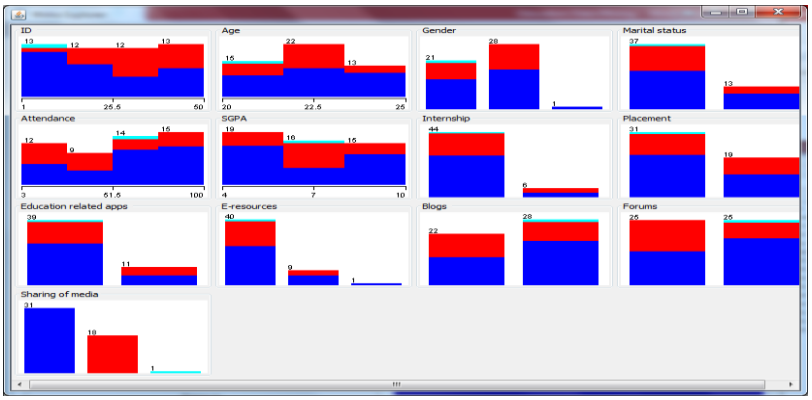

Figure 3(a). *Cluster Analysis*

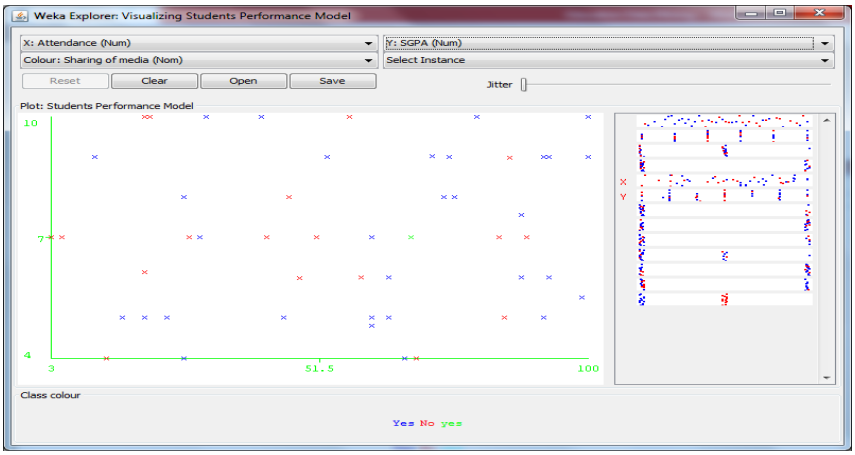

Figure 3(b).

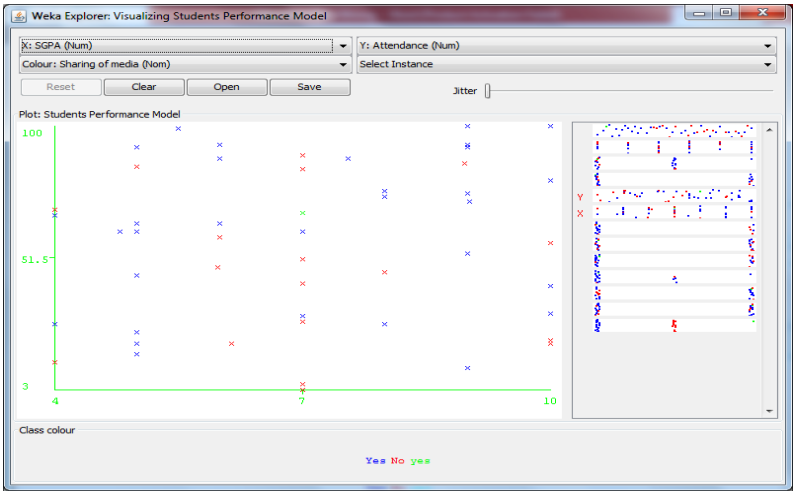

Figure 3(c))

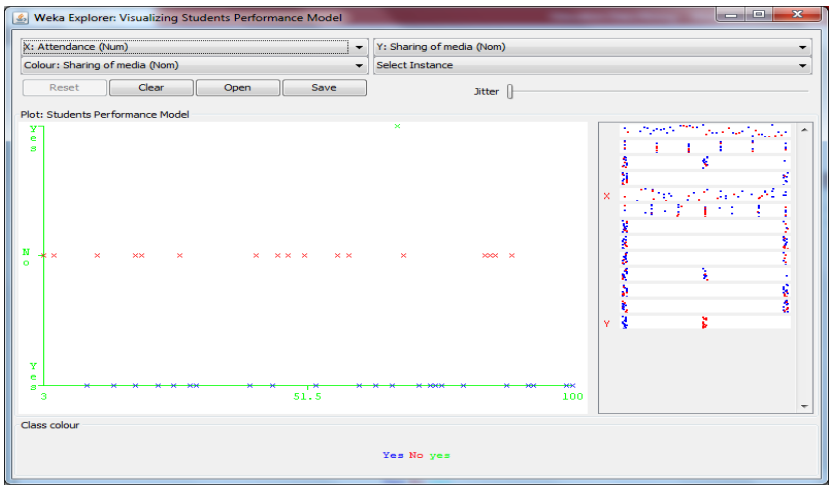

Figure 3(d)

4. Conclusion

Recent advances in the use of technology in educational field create huge data. Data mining can be applied on this data, for understanding and predicting student's performance, and is known as educational data mining. This prediction can be taken into account for improving the effectiveness of teaching-learning process. In this paper, we have developed a model to predict the student's performance by using clustering model with the help of k-means algorithm. It is observed that, the model built using EDM techniques incorporate students overall behavior and pedagogy to analyze the knowledge and teaching and learning outcomes.

References

[1] Goyal, Monika, and Rajan Vohra. "Applications of data mining in higher education." International journal of computer science9.2 (2012): 113.

[2] Patel, MsPriti S., and S. G. Desai. "Various Data Mining Techniques used to Study Student's Academic Performance." (2015).

[3] Bhullar, Manpreet Singh, and Amritpal Kaur. "Use of data mining in education sector." *Proceedings of the World Congress on Engineering and Computer Science*. Vol. 1. 2012.

[4] Romero, Cristóbal, Sebastián Ventura, and Enrique García. "Data mining in course management systems: Moodle case study and tutorial." *Computers & Education* 51.1 (2008): 368-384.

[5] Suman, Pooja Mittal P., and M. Pooja. "A Comparative Study on Role of Data Mining Techniques in Education: A Review." *International Journal of Emerging Trends & Technology in Computer Science* 3.3 (2014): 65-9.

[6] Tair, Mohammed M. Abu, and Alaa M. El-Halees. "Mining educational data to improve students' performance: a case study." *International Journal of Information* 2.2 (2012): 140-146.

[7] Bhise, R. B., S. S. Thorat, and A. K. Supekar. "Importance of data mining in higher education system." *IOSR Journal Of Humanities And Social Science (IOSR-JHSS) ISSN* (2013): 2279-0837.

[8] Romero, Cristóbal, and Sebastián Ventura. "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 (2010): 601-618.

[9] Priya C.,Sana S., "Educational Data Mining: A Critical Study ",BVIMSR'S Journal of Management Research ISSN: 0976-4739.

[10] http://bigdata-madesimple.com/what-is-clustering-in-data-mining/

[11] Berkhin, Pavel. "A survey of clustering data mining techniques." *Grouping multidimensional data* 25 (2006): 71.

[12] https://www.datascience.com/blog/k-means-clustering

[13] Oyelade, O. J., O. O. Oladipupo, and I. C. Obagbuwa. "Application of k Means Clustering algorithm for prediction of Students Academic Performance." *arXiv preprint arXiv:1002.2425* (2010).

[14] https://antibullyingsoftware.com/blog/technology-in-education/11-types-of-education-software-available-to-schools/