International Journal of Scientific Engineering and Science
Volume 3, Issue 2, pp. 16-19, 2019. ISSN (Online): 2456-7361

Use of Data Mining for Evolution of Student Performance

Anil Mishra
Senior Lecturer, Govt. Polytechnic College, Sanawad, Madhya Pradesh, India
Email address: mishra.anil91@gmail.com

***Abstract*—** *In India, the technical education is providing through three layers viz. engineering colleges, polytechnic colleges and ITIs. They are providing degrees, diplomas and certificates.*

Thousands and thousands of students are studying in these institutes. The biggest problems before these institutes are that improving the quality teaching. These institutes have big data describing student performances. In this paper I will make a case study of a polytechnic college that helps in improving the education quality. By the use of data analysis and obtaining the factors that can improve the results of institute. It also increases the chances of success for the students. For all of this I use data mining techniques. I will try to show the significance of data pre-processing for improving the accuracy of the result.

***Keywords*—** *Technical education, Data mining, Knowledge discovery, Data preprocessing, K-Mean Clustering.*

I. INTRODUCTION

In our modern society education is the most essential part of life. Although thousands of men and women in developing and under-developed countries are still not literate. Education does not only mean to read and write but it also has the aims to achieve economic, social, vocational, knowledge, moral and spiritual aims.

India is witnessing the era of science and technology. In our everyday life the control of science and technology is becoming so large that man's existence in this world is quite difficult. This is why, it is necessary to train our people in practically and technically. Technical Education can meet the expanding demands of industries. In India Polytechnics are meant to provide skills after class 10th. The aim of the polytechnic education is to create a middle level connect between technicians and engineers.

To remain competitiveness among the students these institutes need to provide deep and thorough knowledge for a better assessment, evaluation, planning, and decision-making. For this Data mining have many techniques from a variety of fields including databases, statistics, data visualization, machine learning and others. The data mining technology can determine the hidden patterns and associations in these educational data. This can enhance the decision making processes in educational systems. The data mining techniques can extract the patterns like students having similar characteristics,

They can use data mining in finding useful hidden information in the student result database for improving students' learning methodology.

The purpose of this paper is to use data mining techniques for studying students' performance in their stream. For this, we will be using Association rules to compare the student's performance in the subjects at diploma level and will predict the factors. These factors can then be explain students' success or failure.

II. RELATED WORK

In education field, so many researchers have been studied the use of data mining in this area. They discovered the patterns exist in the educational database. Here we will discuss the previous work done by them.

F. Siraj & Abdoulha (2009) has studied data mining techniques like Logistic regression and Decision tree over student data.

C. Romero & S. Ventura (2010) studied the Education system. They described the educational environment. They elaborate various tasks of educational environment that can be solved using different data mining techniques.

S. Anupama Kumar and Vijayalakshmi M.N (2011) worked the decision tree algorithm on internal assessment marks to forecast their performance in the examination. The decision tree forecast the how many students will pass the examination and how many students will fail in the examination.

Hua-long Zhao (2008) has performed Multi-dimensional Cube Analysis (MDCA). They have used OLAP to show that the curriculum selected by any student depends on various factors like teacher, semester etc. They analyze curriculum that helps in making policy for school management, using data warehouse model.

W.M. Tissera et al. (2006) applied many experiment on educational institute. They try to find some relation between subjects. This knowledge helps in making decision for improving education quality.

Qasem A. et al. (2006) find the factors that may influence the student performance in their courses.

S. Ayesha et al. (2010) used K-means clustering for analyzing behavior of learning of students.

Hongjie Sun (2010) make a research on student learning using data mining. The researches aimed at result evaluation and apply it for improving learning skills.

International Journal of Scientific Engineering and Science
Volume 3, Issue 2, pp. 16-19, 2019. ISSN (Online): 2456-7361

III. KNOWLEDGE DISCOVERY PROCESS

There are many data mining techniques that uses huge data to discover unknown patterns and associations which are helpful in decision making. The steps for extracting knowledge from data are as shown below:

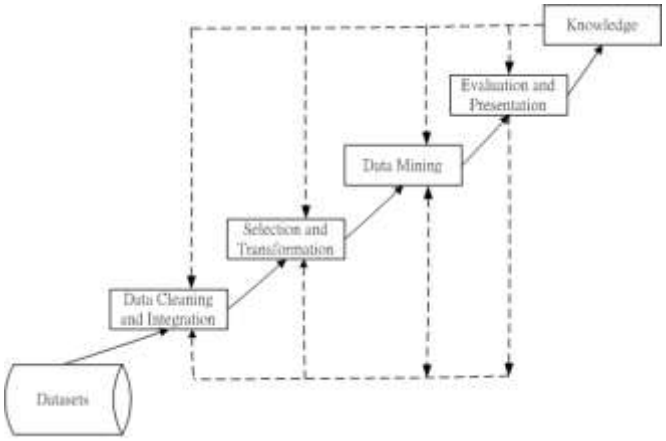

Fig. 1. Steps for extracting knowledge from data.

3.1 Cluster Analysis and K-Mean Algorithm

Cluster analysis or clustering is a type of data mining technique in which the items are taken in a group based on the similar characteristics. In this techniques there are no predefined class-level.

K-Mean Clustering is an algorithm to group items depending on their characteristics into K number of groups. The major task is to label K centroid, one centroid for each group. This is a technique that tries to minimize the sum of squares of distances between data and the corresponding cluster centroid.

Thus the objective of this algorithm is to minimize the squared error function, which is given by:

$$J(D) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|a_i - d_j\|)^2$$

Here, $\|a_i - d_j\|$ is the Euclidean distance between a_i and v_j.
c_i is the number of items in i^{th} cluster.
c is the number of cluster centroid.

Algorithmic steps for k-means clustering

Let X = $\{a_1,a_2,a_3,\ldots\ldots,a_n\}$ be the set of data items and D = $\{d_1,d_2,\ldots\ldots,d_c\}$ be the set of centroid.

1. Arbitrarily Choose *'c'* cluster centroid.
2. Calculate the distance between each data item and cluster centroid.
3. Allocate the data item to the cluster centroid whose distance from the cluster centroid is smallest of all the cluster centroid.
4. Recalculate the new cluster centroid using:

$$d_i = (\frac{1}{c_i}) \sum_{j=1}^{c_i} a_i$$

where, *'c_i'* represents the number of data items in *i^{th}* cluster.

5. Recalculate the distance between each latest cluster centroids and data items.
6. If no data item was reallocated then stop, otherwise repeat from step (3).

3.2 Data Collection

The sample data is taken from a polytechnic college of Madhya Pradesh. The data contains the result of students of their diploma examination. The data has Roll Number and Programme Code i.e. stream, Roll Number and Marks for each student.

3.3 Data Cleaning

The collected data require to be clean and transformation.

We clean the student data by removing records of all the students who are not from the stream Computer Science and Engineering stream.

3.4 Data Transformation

In our data those students who got marks less than 33; their marks entered in their MARKS column, * appended with marks like 04 *.

Also those students who passed the subject with grace mark G is added with their marks like 28 G.

Since these attribute is now identified with Character type, so we transform these attribute by removing * sign or G character with marks. We put only 28 instead of 28 G or only 6 instead of 06 *. By doing this the MARKS attribute is now identified as Numeric type.

The highlighted records need to be transformed:

PROG_CD	ROLL_NO	SUBJECT_CD	MARKS
C04	'08027C04002	601	64
C04	'08027C04002	602	49
C04	'08027C04002	603	52
C04	'08027C04002	613	49
C04	'08027C04002	623	35
C04	'08027C04004	601	52
C04	'08027C04004	602	42
C04	**'08027C04004**	**603**	**21 ***
C04	'08027C04004	613	41
C04	'08027C04004	623	33
C04	**'08027C04004**	**101**	**26 ***
C04	**'08027C04004**	**401**	**18 ***
C04	'08027C04004	504	33
C04	'08027C04006	601	60
C04	'08027C04006	602	42
C04	'08027C04006	603	46
C04	'08027C04006	613	33
C04	**'08027C04006**	**623**	**28 G**
C04	**'08027C04007**	**303**	**06 ***
C04	**'08027C04007**	**304**	**07 ***
C04	**'08027C04007**	**305**	**18 ***
C04	**'08027C04007**	**306**	**04 ***
C04	**'08027C04007**	**401**	**05 ***
C04	'08027C04007	402	23
C04	'08027C04007	403	33

Fig. 2. Data table before transformation.

IV. ANALYZING CLUSTERING

In this study the data mining tool Tanagra is used to analyze the clustering technique. A part of data table that is used with this tool rules is shown here:

International Journal of Scientific Engineering and Science
Volume 3, Issue 2, pp. 16-19, 2019. ISSN (Online): 2456-7361

Fig. 3. Version of data mining tool Tanagra used.

PROG_CD	ROLL_NO	SUBJECT_CD	MARKS
C04	'0627182	508	60
C04	'0627188	204	7
C04	'0627188	105	8
C04	'0627189	105	33
C04	'0627194	401	39
C04	'0627194	201	5
C04	'0727182	508	51
C04	'0727186	105	4
C04	'0727186	204	15
C04	'0727186	101	24
C04	'0727190	105	37
C04	'0727191	106	33
C04	'0727191	508	40
C04	'0727191	401	50
C04	'08019C04030	613	36
C04	'08019C04030	623	37
C04	'08019C04030	504	37
C04	'08019C04030	602	39
C04	'08019C04030	601	61
C04	'08019C04030	603	61
C04	'08019C04030	201	0
C04	'08019C04030	401	19
C04	'08027C04002	623	35
C04	'08027C04002	602	49

Fig. 4. Data table used with Tanagra tool.

Below are the snapshots of Tanagra showing activities at different stages with K_Mean Clustering Process.

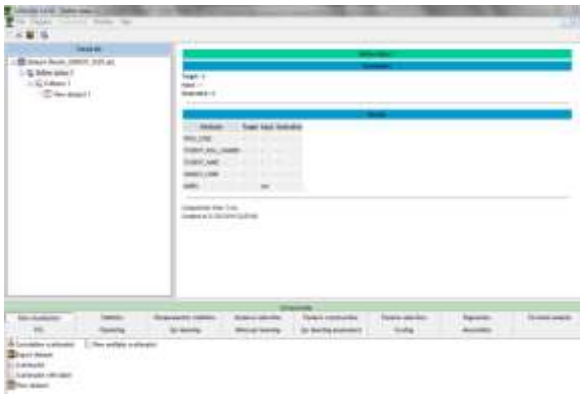
Fig. 5. Data table with different attributes.

4.1 Analysis of Clustering Technique

After analyzing the generated K-Mean Clustering technique it is observed that the students who have scored badly in their subjects are grouped with Cluster_ID#3 and the student those performed good in their examination grouped with Cluster_ID#1. The average performance is grouped with Cluster_ID#2 as shown in figure 8.

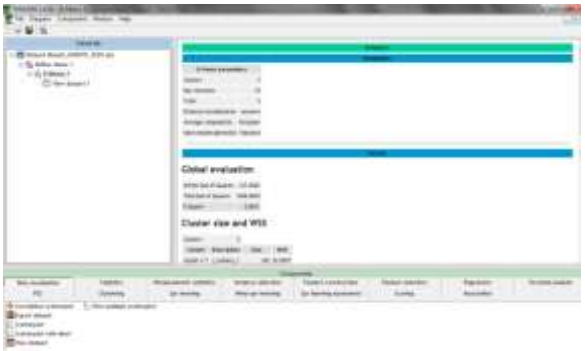
Fig. 6. K-Mean with cluster size.


Fig. 7. K-Mean with clusters Vs input attributes and Cluster Centroids.

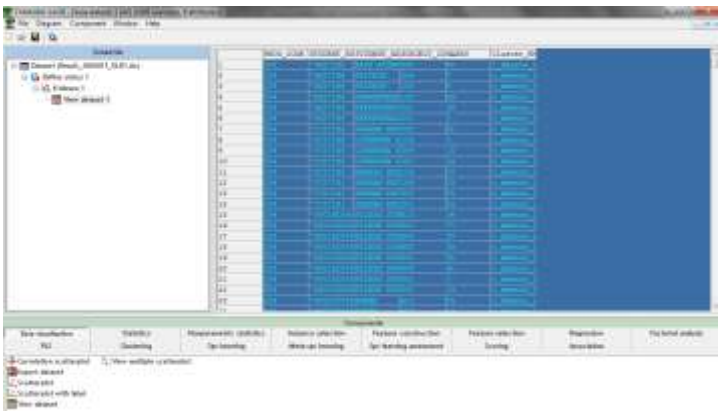
Fig. 8. View of data set with Cluster_ID.

The generated k-mean data view are very helpful to college administration and directorate level. They may use this unknown information and patterns discovered in the future planning for the betterment of the study. It would not only benefit students but the academic institute also.

V. CONCLUSION

The paper analyzed the prospective application of one of the data mining technique K-Mean Clustering in improving the worth of students' performances at diploma level.

REFERENCES

[1] F Siraj and Abdoulha, "Uncovering hidden information within University's student enrollment data using data mining", *Third Asia International Conference on Modelling and Simulation*, 2009.
[2] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 6, November 2010.

[3] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar and M. Inayat Khan, "Data mining model for higher education system", *Europen Journal of Scientific Research*, vol. 43, no. 1, pp. 24-29, 2010.
[4] W. M. Tissera, R. I. Athauda, and H. C. Fernando "Discovery of strongly related subjects in the undergraduate syllabi using data mining", *IEEE International Conference on Information Acquisition*, 2006.
[5] Hongjie Sun, "Research on student learning result system based on data mining", *IJCSNS International Journal of Computer Science and Network Security*, vol. 10, no. 4, April 2010.
[6] Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, Mustafa I. Al-Najjar, "Mining student data using decision trees", *ACIT' 2006: The International Arab Conference on Information Technology*.
[7] J. Abonyi and B. Feil, *Cluster Analysis for Data Mining and System Identification*, Boston, MA: Birkhäuser Basel, 2007.
[8] M. S. Aldenderfer and R. K. Blashfield, *Cluster analysis*. Newbury Park, CA: Sage Publications, 1984.
[9] M. R. Anderberg, *Cluster analysis for applications*, New York: Academic Press, 1973.