Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 1, January -2015

Mining Student Data to characterize similar behavior Groups using Distributed Data mining For Performance Improvement

Pooja Bhatt[1], Dinesh Vaghela[2],Dr.Priyanka Sharma[3]

[1]*Computer Engg. Department, Parul Institute of Technology Pooja.bhatt207@gmail.com*
[2]*Computer Engg. Department, Parul Institute of Technology, dineshcsepit@gmail.com*
[3]*Computer Engg. Department, Parul Institute of Technology pspriyanka@yahoo.com*

Abstract: *Educational data mining is very important for characterize similar behavior groups for the student's performance Improvement. In this paper, we have provide the proposed flow that how to characterize similar behavior groups using distributed data mining. There are preprocessing the data, data mining techniques to discover association and collection of the attributes, classification, and clustering techniques. In these tasks, we extracted knowledge that describes students' behavior.*

Keywords: *Educational data mining, Distributed data mining*

I. INTRODUCTION

Educational Data Mining mainly focuses on developing new tools and algorithms for discovering data patterns [1]. The ability to identify is very important in educational environments for providing student‟s academic performance. The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data [1]. The main functions of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [2]. Data mining and knowledge discovery applications have a rich focus due to its significance in decision making. Data mining techniques have been introduced into new fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc. In this paper, there are provided some procedures based on the observation and some suggestions those can used to make decision as to improve the student academic performances.

The main objective of this paper is to use data mining methodologies to study students‟ performance in the courses. Data mining provides many tasks that could be used to study the student performance. Information like Attendance, Seminar and Assignment marks were collected from the student‟s management system and student‟s database, to characterize similar behavior groups of student data which effects the performance of students[4].There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments [3]. Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others. Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering.

Distributed data mining is originated from the need of mining over decentralized data sources and data from different sites. The field of Distributed Data Mining (DDM) deals with these challenges in analyzing distributed data It offers many algorithmic solutions to perform different data analysis and mining operations in a fundamentally distributed manner that pays careful attention to the resource constraints. In this paper is a survey concerned with Distributed Data Mining algorithms, methods and trends in order to discover knowledge from distributed data in efficient way[5]for the characterize students based on the student‟s performance .

In this paper, through the clustering groups of students data and attributes which affect the student‟s performance will help to know where the improvement can be done.

From that student‟s performance techniques will be provided.

II. SCOPE

In this paper, with the help of educational data mining Student‟s performance attributes will be clustered and then depends on clusters it defined groups of performance of student‟sbehavior.

From the clusters Student‟s performance are identified with the help of that characterization of the student‟s data. The results reveal that the student‟s performance level can be improved in university result by identifying students who perform poorly in unit Test, Attendance, Assignment and giving them additional guidance to improve the university result.

III. MOTIVATION

Many independent research works have been undertaken using the data mining techniques to work on education databases and other educational institutions. The main purpose of data mining on such educational databases is to create meaningful learning outcomes, planning the academic levels for the students and intervening as also to predict the behavior and research of the alumni. Keeping this as a base, there is a way proposed for understanding the students‟ general opinion about their satisfactions and discontentment in the educational methodologies and to also predict their specific interests in the fields of study. Security, Filer control can easily cured by distributed data mining. Time complexity of the distributed data mining is reducing than the central data mining process. Distributed data mining process is more accurate and efficient than the central data mining process.
Based on these interests, their current education, specific areas of failures can be accurately detected and they can be made job ready for the market. The method evaluates the students‟ performance based on parameters like attendance, oral examinations, mid and final exams and conveys that information to the teachers in charge.
This can help the teachers evaluate their students‟ performance and help in reducing the attrition rate.

IV. RELATED WORK

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Mining in educational environment is called Educational Data Mining.

AhmadN [6] did the research on Performance of off-campus engineering students in a Malaysian University.

The survey was categorized into three parts; the students' academic performance, their attitude and class behavior and their soft skills.

Results indicate that the student's mathematical skills and knowledge in engineering fundamental basics are very low. In addition, their attitude and class behavior are probably the reason why they cannot score high CGPA.
SuchitaBorkar [1] did Attributes Selection for

Predicting Students‟ Academic Performance using Education Data Mining and Artificial Neural Network and selected effective attributes for student‟s performance.

EDM using association rule mining algorithm and artificial neural network in predicting students‟ performances in university result

Hoe[7]has Analyzed students records to identify patterns of students' performance through the data mining techniques.

The data mining technique used to identify the significant variables that affects and influences the performance of undergraduate students.

Students „demographic and past academic performance data are then used to study the academic pattern.

Bhavesh [2] has studied and Evaluated of Student Performance Parameters using statistical features of excel. Demographic information and behavioral data of students are used to get the result that which parameters are highly affected on student‟s SPI result. The raw data was pre-processed in terms of filling up missing values, transforming values in one form into another and relevant attribute/ variable selection.

J. James Manoharans[8] has Discovered Students‟

Academic Performance Based on GPA using K-Means Clustering Algorithm.

The k- mean clustering algorithm combined with deterministic model to analyse and monitor the student‟s results and their performance. By this k-mean clustering we can get more efficiency on monitoring the progress of academic performance of students in higher Institution to provide accurate results in a short period of time. In this paper, we applied the methodology to find out the various interesting pattern by taking the student test scores.

V. PROBLEM DEFINITION

- To analyse student‟s performance and providing suggestion for improvement of student‟s academic Performance is the major challenge.
- With the help of literature survey, till now all clustering algorithms are performed for student‟s performance analysis in local environment.
- In our work, Clustering algorithm will be applied on Distributed environment.
- Improving accuracy and efficiency in distributed environment is problem definition.

VI. PROPOSED WORK

A. DATA COLLECTION

In our case study we collected the students data from data based management system course held at the Parul Institute of Technology. The sources of collected data were: personal records and academic records of students, course records and data came from e-learning system. Then, we got information about how much student benefited from resources, such as using ebooks, research papers and old exams available on the system. Also, we got the results of students' grades in solving exercises available in the system.

B. DATA SELECTION AND TRANSFORMATION

In this step only those fields were selected which were required for data mining [9]. A few derived variables were selected. While some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from the database are given in Table I for reference.

Table 1: Selected Attributes for Characterizing students data

Attributes	Description	Possible Values
PM	Previous Semester Marks	First –>60%, Second –>45% and <60%, Third –>36% and < 45%, Fail < 40%.
AT	Attendance	{Poor , Average, Good}
AS	Assignment Submission	{Yes , No}
LP	Lab Performance	{Yes, No}

SP	Seminar Performance	{Poor , Average, Good}
EM	End Semester Marks	First – >60%, Second – >45% and <60%, Third – >36% and < 45%, Fail < 40%.
GP	General Proficiency	{Yes, No}
MM	Mid Semester Marks	{Poor , Average, Good}

The domain values for some of the variables were defined for the present investigation as follows:

PM: PM means previous semester marks in B.Tech course. It is split into four class values: First > 60%, Second >50 &<60%, Third >40 &<50% and Fail < 40%.

MM: Mid Semester Marks: mid exams conducted internally in college. In each semester, two written exams and two online exams are conducted and averages of four tests are used to calculate internal marks. It is split into three class values: Poor, Average and Good

SP: Seminar performance participated outside the campus. Seminar performance is evaluated into three classes: Poor – Presentation and communication skill is low, Average – Either presentation is fine or Communication skill is fines, Good – Both presentation and Communication skill is fine.

AT: Attendance of student. To attend the university exams, each and every student must have minimum 75% attendance. Attendance is divided into three classes: poor < 60%, average 60% >&< 75%and good >75%.

AS: Assignment Submission Assignment Submission is divided into two classes :Yes – student submitted assignment, No – Student not submitted assignment.

LP: Lab Performance LP means Entire Lab Work. Entire lab work means both internal and external lab work and exams. Entire Lab work is divided into two classes: Yes – student completed lab work, No – student not completed lab work .

EM :End semester Marks obtained in BE semester and it is declared as response variable. It is split into five class values: First – >60% , Second – >45% and <60%, Third – >36% and < 45%, Fail < 40%.

GP: General Proficiency performance. Like seminar, in each semester general proficiency tests are organized. General Proficiency test is divided into two classes: Yes – student participated in general proficiency, No – Student not participated in general proficiency.

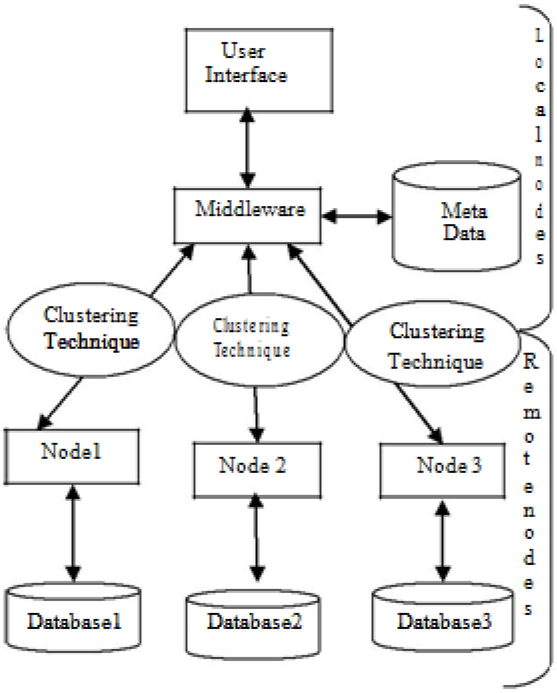
C. PROPOSED FRAMEWORK:

Following steps of the method will be applied to the System:

- In the first step, Clustering algorithm will be applied to the local nodes are showed in figure1 which are node1, node2 and node3.

All nodes will make the clusters as per similar behavior groups of the database they obtain which are database1, database2, database3.

- Individual Clusters created by individual nodes will be combined with all remote nodes .



Steps for Proposed Work:

The steps for proposed work are following:

1. End user will enter the requested data on the User Interface layer like Student id.
2. Then, middleware will generate request according user‟s request.
3. All nodes which are node1, node2 and node3 have individual databases and clustering algorithm will be applied on every node.
4. Clusters are created by every node will send to the middleware.
5. Middleware will have the metadata so it will consolidate all clusters sent by nodes and will make a global clusters from local clusters.
6. Middleware will contain all servers id means address of local nodes.

7. Middleware will send the reply to the end user by this method.

R=n1R1+n2R2+n2R3

Where R= Final Result

n1R1= Result of node1 or server1 n2R2= Result of node2 or server2 n3R3= Result of node3 or server3

8. In the end, final result will be sent to the User interface.

VII. CONCLUSION AND FUTURE WORK

In this paper, the clustering task is used on student database to characterize similar behaviour groups of the students division on the basis of previous database. As there are many approaches that are used for data clustering. Information‟s like

Attendance, Lab work, Seminar and Assignment performance were collected from the student‟s previous database, to analyse the performance at the end of the semester. This study will help to the students and the teachers to improve the division of the student. In the Comparison of the central data mining clustering techniques, the Distributed data mining is more efficient, scalable and performance is better than the central data mining techniques. This study will help to the students and the teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination. This study will help to students and teachers how they can improve the performance of the students.

For the future work, many improvements in efficiency of the clustering algorithm with rich and different kind of dataset can use for the enhancement.

REFERENCES

[1] SuchitaBorkar, K.Rajeswari "Attributes Selection for Predicting Students‟ Academic Performance using Education Data Mining andArtificial Neural Network" International Journal of Computer Applications (0975 – 8887) Volume 86 – No 10, January 2014

[2] Mr.Bhavesh Patel, Mr. Bharat Prajapati, Dr.JyotindraDharwa, Dr. A. R. Patel "Study and Evaluation of Student Performance Parameters using statistical features of excel" International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-2, Issue2, March-2014

[3] PimpaCheewaprakobkit "Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program" Proceedings of the International Multi Conference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong

[4] Brijesh Kumar Bhardwaj,Saurabh Pal "Mining Educational Data toAnalyze Students ‟Performance (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

[5] Josenildo C. da Silva2, Chris Giannella1, Ruchita Bhargava3 " Distributed Data Mining and Agents"

[6] AhmadN,Yahya,R.Salam,K;Buniyamin,N"Performance of off-campus engineering students in a alaysian University: An investigation"

[7] Hoe, A.C.K ; Ahmad, M.S; Tan Chin Hooi; Shanmugam, M. ; Gunasekaran, S.S. ; Cob, Z.C. ; Ramasamy"Analyzing students records to identify patterns of students' performance"

[8] . James Manoharans,Dr. S. HariGanesh,M.LovelinPonnFelciah, A.K. ShafreenBanu"Discovering"Discovering Students‟ Academic Performance Based on GPA using K-Means Clustering Algorithm"IEEE2014

[9] P. Ajith, B. Tejaswi, M.S.S.Sai"Rule Mining Framework for Students Performance Evaluation"International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013