Data Clustering in Education for Students

Shashikant Pradip Borgavakar, Mr. Amit Shrivastava, Prof. Preetesh Purohit

[1]*Research Scholar, Computer Science & Engineering Department, Swami Vivekanand College of Engineering Indore, India*
[2]*Asst. Professor, Computer Science & Engineering Department, Swami Vivekanand College of Engineering Indore, India*
[3]*Professor, Head of Computer Science and Engineering Department, Swami Vivekananda College of Engineering, Indore, India*

Abstract –*Data clustering used to maintain significant relationships for large data set or warehouse by using extraction of data. We are analyzing student behavior by using data clustering technique means K-Means Clustering. Class mid and final exam, assignment, credit test by using this factor we are evaluating studied. To evaluate accurate result we are required to correct student mark detail and it was strongly recommended that to get correct result we are required correct input data. This study will help professor to reduce the failed ratio to significant level and improve the performance of students.*

Key Words: K-means, Database, Student evaluation etc.

1. INTRODUCTION

Data clustering is process of extracting large data which are valid, hidden pattern and useful data. Day to day data which are store for education is increasing rapidly. For future prediction clustering technique is most significant technique. Here we are partitioning homogeneous groups with their characteristics and performance to evaluate result. This application can help both instructor and student to enhance the education quality. This study makes large data set of students into groups of student by suing their characteristics.

2. LITERATURE SURVEY

Irjet Template sample paragraph .Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Table -1: Sample Table format

Research Paper	Improving the Accuracy and Efficiency of the k-means Clustering Algorithm	An Iterative Improved k-means Clustering	Refining Initial Points for K-Means Clustering	Comparison of various clustering algorithms
Problem being addressed	Lower accuracy and efficiency	Number of Iterations are Less	Estimate is fairly unstable due to elements of the tails appearing in the sample	Which clustering algorithm is best
Importance of the problem	algorithm requires a time complexity	Total number of iterations required by k-means and improved k-means is much larger	Importance of the problem of having a good initial points	Way of Process
Gap in the prior work	Accuracy and Efficiency is most complicated to reducing	Check multiple iterations	To finding Initial Points	Finding algorithm
Specific research questions or research objective	To Overcome the problem of Accuracy and Efficiency	This paper presented iterative improved k-means clustering algorithm that makes the k-means more efficient and produce good quality clusters	A fast and efficient algorithm for refining an initial starting point for a general class of clustering algorithms has been presented	data mining is that to discover the data and patterns and store it in an understandable form
Broad outline of how the author solved the problem	Using K-Means clustering Algorithm and The enhanced Method	Iteration improve k-means cluster algorithm	Using Clustering Cluster	Applied DBSCAN and OPTICS algorithms
Details of implementation of procedure	Phase 1 of the enhanced algorithm requires a time complexity of O(n2) for finding the initial centroids, as the maximum time required here is for computing	Dividing number of parts then calculate centers and decide membership of patterns then repeat same steps	Results on Real Word Data	All clustering algorithm process and find

	the distances between each data point and all other data-points in the set			
Key contribution of the paper claimed by the author.	define k centroids, one for each cluster	iterative improved k-means clustering algorithm	Clustering Clusters	K-Means clustering Algorithm

3. DATA CLUSTERING

Data Clustering is a process which partitions a given data set into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects are in different groups. It is the most important unsupervised learning problem. It deals with finding structure in a collection of unlabeled data. For better understanding please refer to Fig I.

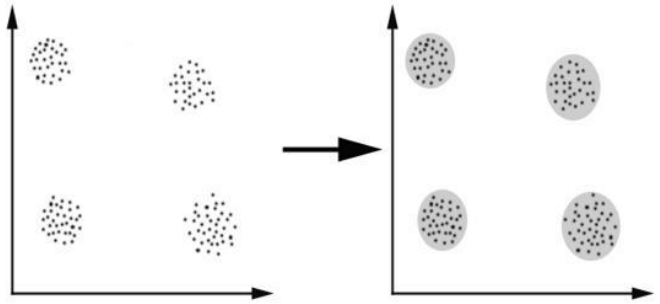


Fig.1: 4 clusters formed from the set of unlabeled data

4. CLUSTERING IN HIGHER EDUCATION

Education is very important factor in our life. Education make perfect in their professional work and it is very important to be part of educated society to grow in career. To make stable country every individual has to become professionally educated, it will better for country. We use clustering in education to classify students in their academic performance. This study helps system management to grow their academic performance and make better education system. This data clustering methodology helps system to fill gaps in higher education.

5. PROPOSED MODEL

In College academic performance are calculated by credit test, internal and external exam test. Internal mark are class test, practical's marks, lab test, quiz and attendance. External tests are semester test and final exam. By using internal exam test and previous grade we are calculate and predict final grade of student.

1. If prev-grade=high, quiz=good, assignment=complete, lab-performance=good ,class-test=good, attendance=regular and then final-grade=good
2. If prev-grade=average, quiz=good, assignment=incomplete lab-performance=good Class-test=average and attendance=regular then final-grade= average
3. If prev-grade=low, quiz=average, assignment=incomplete, lab-performance= poor mid-term=low and attendance=irregular then final-grade=low.

This study tries to identify weak student before final exam and save them from failure. Professor can make right step against this weak students to increase their performance in final exam.

6. K-MEANS CLUSTERING ALGORITHM

This technique is very old one. By using K means we are processing large data set and get valuable information from this large data set. Pseudo code is given below:

Step 1. Accept the number of clusters to group data into and the dataset to cluster as input values.

Step 2. Initialize the first K clusters - Take first k instances or - Take Random sampling of k elements.

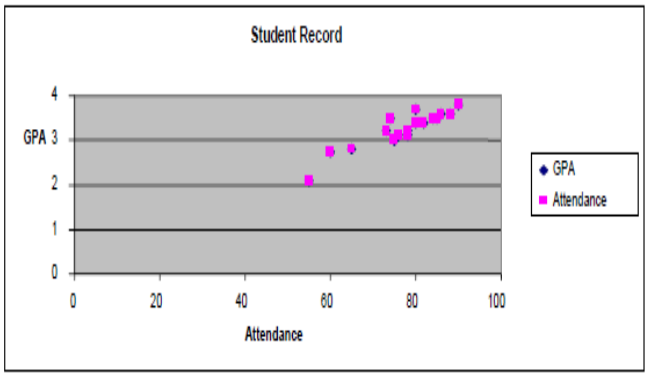
Step 3. Calculate the arithmetic means of each cluster formed in the dataset.

Step 4. K-means assigns each record in the dataset to only one of the initial clusters - Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).

Step 5. K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

7. RESULT AND DISCUSSION

The study produced following results:

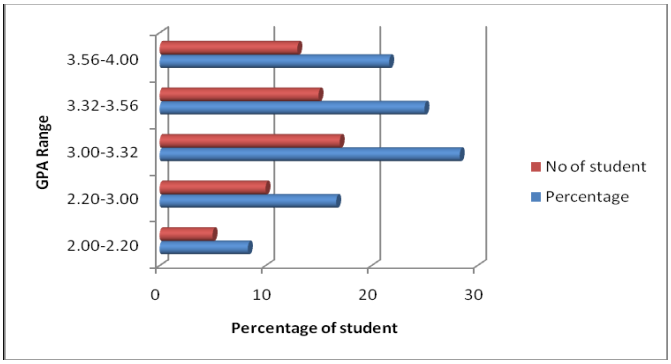


Graph.1: Relationship between GPA and Attendance ratio

We grouped the students regarding their final grades in several ways 3 of which are: • Assign possible labels that are same as number of possible grades. • Group the students in three classes "High", "Medium" and "Low". • Categorized the students with one of two class labels "Passed" for grade above 2.20 and "Failed" for grade less than or equal to 2.20

Class	GPA	No of student	Percentage
1	2.00-2.20	5	8.33
2	2.20-3.00	10	16.67
3	3.00-3.32	17	28.33
4	3.32-3.56	15	25
5	3.56-4.0	13	21.67

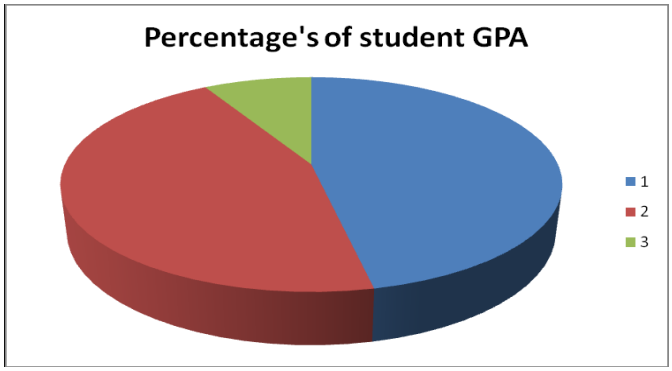
Here, I cluster student among their GPA that means, from GPA 2.00- 2.20 we have 8.33% student. From 2.20-3.00 student percentage is 16.67%. From 3.00-3.32 we have 28.33%. From 3.32-3.56 percentage is 25% .The percentage is 21.67% between GPA 3.56-4.00. The graphical representation of GPA and the percentage of student's among the student are given below.



Graph 2: Number and percentage of students

Class	GPA	No of student	Percentage
High	>=3.50	28	46.67
Medium	2.20<=GPA<	27	45
Low	3.5 <=2.20	5	8.33

After applying clustering onto the student date set, we group the student into three categories. First is High, second is Medium, and the last one is Low. Graphical representation of these three categories is given below:



Graph 3: Shows the percentage of students getting high, medium and low GPA

REFERENCES

[1] Alaa el-Halees (2009) Mining Students Data to Analyze e-Learning Behavior: A Case Study.

[2] Behrouz.et.al., (2003) Predicting Student Performance: An Application of Data Mining Methods With The Educational Web-Based System Lon-CAPA © 2003 IEEE, Boulder, CO.

[3] Connolly T., C. Begg and A. Strachan (1999) Database Systems: A Practical Approach to Design, Implementation, and Management (3rd Ed.). Harlow: Addison-Wesley.687

[4] Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57)

[5] Galit.et.al (2007)Examining online learning processes based on log files analysis: a case study. Research, Refelection and Innovations in Integrating ICT in Education.

[6] Henrik (2001) Clustering as a Data Mining Method in a Web-based System for Thoracic Surgery: © 2001

[7] Han,J. and Kamber, M., (2006) "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

[8] Kifaya(2009) Mining student evaluation using associative classification and clustering. Communications of the IBIMA vol. 11 IISN 1943-7765. ZhaoHui. Maclennan.J, (2005). Data Mining with SQL Server 2005 Wihely Publishing, Inc