Evaluation of Student Performance using Data Mining over a Given Data Space

OTOBO Firstman Noah, BAAH Barida, TAYLOR Onate Egerton

Abstract*— The volume of data generated every year in our institutions is enormous, due to this large volume of data there is the need to provide a efficient system support to aid in good decision making process; this is what necessitated this research paper which is all about the evaluation of student performance using data mining technique over a given data space. In this paper we are going to look at the various data mining techniques, data mining algorithms and the k-means clustering technique. In this paper, the performance evaluation of students, were presented using data mining technique and cluster checking. The system examined students who gained admission into the University of Port-Harcourt through the University Matriculation Examination (UME) and through Basic studies programme with the aim of finding out variations in their performance when they graduate from the university. The evaluation was done using data mining technique to find out the ratio that falls into grouping of the grading in the various classes using the cumulative grade point average (CGPA) and the students who failed out. The system was able to cluster, analyze and report the relative performance of each of the groups of the students used in the research work. Finally, the system was implemented using Apache, MySQL, PHP, internet explorer, NetBeans IDE 6.8 and XAMPP web server.

***Index Terms*— Data Mining, Data Space, Clustering, Database**

I. INTRODUCTION

In the light of the increased possibilities in modern society for institution, data volume increases every day, to gather data cheaply and efficiently, more efficient database solutions are needed to support temporal and logical consistency over the large volume of data. Universities are facing the immense and quick growth of the volume of educational data [1]. Databases are distributed across more than one location and it is connected by communication link in which each site is a database system in its own right, but the sites have agreed to work together so that a user at any site or node can access data anywhere in the network if the need arises, and it acts as if the data were stored at the user own site.

The process of data mining over distributed databases is to efficiently extract or "mine" knowledge from large amount of data that is distributed over multiple sites. There is one important reason to choose DDM, the reason is that data is become very large, it's hard to store it at a single site, or it's inefficient or incapable to mine such large data at a single site.

In such a case, the data may be decomposed into some parts and it is distributed at different site then the mining operations can be performed for each of the site, then the results can then be combined to gain global result.

Manuscript received on September, 2013.

OTOBO Firstman Noah, Msc Computer Science, University of Port Harcourt, Port Harcourt, Nigeria.

BAAH Barida, Msc Computer Science, University of Port Harcourt, Port Harcourt, Nigeria.

TAYLOR Onate Egerton, Computer Science Dept., University of Science and Technology, Port Harcourt, Nigeria.

This mining is easy, efficient, and flexible. Easy means the system is easy to use and efficient mean the system can mine the knowledge from data with high performance, and flexible means the system can deal with very large data distributed across many nodes. Data mining process involve the analysis of data from different perspective and summarizing it into useful information. The data to mine is based on students' data set in the university of Port-Harcourt, students that where admitted through BASIC(Basic Studies) and UME(University Matriculation Exam) and to check their performances, this will enable the institution to know the set students to admit either through BASIC or UME.

II. RESEARCH METHODOLOGY

A detail feasibility study was carried out in this research paper; the method that was used to arrive at the study is objective review of performance data. The university maintains information about students record, this includes data about time and results of the students, and the number of students that applied for admission, and the number of students that where admitted into various departments in the University.

The list of students admitted through BASIC and UME where collected from the admissions office, Delta Pack University of Port-Harcourt and the result of the students were collected from the department of computer science University of Port-Harcourt. Student record system holds information about student records, for example student examination results, assessment, name, year of enrollment, it is the most important data source for this work. The data of the students are capture and stored in a relational formant, once we have identify the characteristic of students who are unsuccessful in the past semester and year and to partition students into homogeneous group according to their performance. We can compare past performance with current performance at different levels.

III. DATA MINING

Data Mining is the process of extracting pattern or useful data from large database or is the process of analyzing data from different perspective and summarizing it into useful information. [2] define data mining as the "the process of discovering implicit patterns in data stored in a data warehouse and using those pattern for business advantage" while(Berzal *et al)* define Data mining "as a generic term which covers research results, techniques and tools used to extract useful information from large data base". Data mining is not the same thing as data warehousing or data analysis. Data mining is a dynamic process that enables a more intelligent use of data warehouse than data analysis [4]. The aim of data mining is to examine the database for regularities that may lead to a better understanding of the domain described by the database.

In data mining we generally assume that the database consists of a collection of individuals. Depending on the domain, individuals can be anything from customers of a bank, molecular compound, or books in the library. For each individual, the database gives us detailed information concerning the different characteristics of the individual, such as the name and address of customers of a bank or the account owned.

A. Pre Processing

Before Data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined in an acceptable time frame. A common source of the data is a database or data warehouse. Pre-process is essential to analyze the multivariate data sets before clustering or data mining

B. The Knowledge Discovering Process

Most authors have different definition for data mining and knowledge discovering. (Goebel and Gruenwald,1999) define knowledge discovery in databases (KDD) as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" and Data mining as "the extraction of patterns or models from observed data." [5]. Berzal *et al.* define KDD as "the non-trivial extraction of potentially useful information from a large volume of data where the information is implicit (although previously unknown)."

The knowledge discovering and data mining (KDD) process can roughly be separated into the following steps;

1. **Data Identification**; the target subsets of data and the attributes of data of interest are identified by examining the entire raw data set
2. **Data selection**; select the data set to be studied
3. **Data cleaning**; clean the data of duplicates and error, field values are transformed to common units and some new fields are created by combing exiting fields to facilitate analysis. The data is typically put into relational format.
4. **Data mining**, data mining algorithms are applied to extract interesting pattern
5. **Evaluation**; the patterns are presented to end-user in an understandable form. E.g. for example, through visualization.

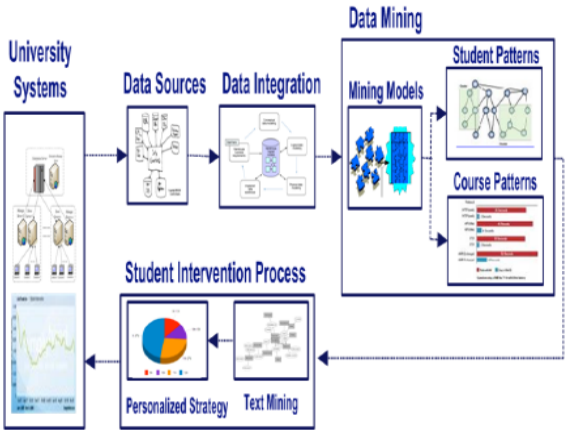

Fig. 1: Diagram showing Data Mining Process of the University

IV. DATA MINING TECHNIQUES

The data mining techniques allow user to analyze data from many different dimensions or angles, categorize it, and summarize the relationship identified. Data mining techniques provide the algorithms that fuel the KDD process. Data mining is the essence of the KDD process. If data mining is being discussed it is observed that the process of KDD is being used. The major opportunities for improvement in data mining technology are scalability and accuracy of data mining techniques. Data mining techniques such as Neutral networks, Genetic algorithms, Regression, Statistical analysis, Machine learning, Decision tree, Cluster analysis are prevalent in the literature on data mining. The SQL/MM: Data Mining extension of the SQL: 1999 support data mining models such as frequent itemsets and association rules, clusters of records, regression trees, and classification tree.

V. DATA MINING ALGORITHMS

There are several types of data mining algorithm that an analyst can apply for well-chosen data sets. Most of the data mining algorithms that are employing in a specific task are briefly discussed below:

1. Counting Co-Occurrences

The problem of counting co-occurring items which is motivated by problems such as market basket analysis. A market basket is a collection of items purchased by a customer in a single customer transaction. A customer transaction consists of a single visit to a store, a single order through a mail-order catalog or an order at a store on the web. A common goal for retailers is to identify items that are purchased together. This information can be used to improve the layout of goods in a store or the layout of catalog pages [6].

2. Frequent Itemset

The frequent itemset algorithms is an algorithms used to scan a relation to determine items that frequently occur after several iteration at different levels. An itemset is a set of item. The support of an itemset is the fraction of transactions in the database that contain the entire item in the itemset. Frequent itemset is the set of item that occur mostly in a relation. A records that is stored into group and showing all tuples in a group having the same attributes, it may be observed that most of the item appear mostly. We have observe that there is redundancy in the relation: it can be decomposed storing these attribute into separate table and stored creating such denormalized table for ease of data mining is commonly done in the cleaning step of the KDD process. A more sophisticated algorithm is the iterative generation and testing of candidate itemsets

Generating candidate itemset by adding an item to a known frequent itemset is to limit the number of candidate itemset using the a priori property. [6],[7].

The priori property implies that a candidate itemset can be frequent only if all its subsets are frequent. Thus we can reduce the number of candidate itemset further- a priori, or before scanning the relation.

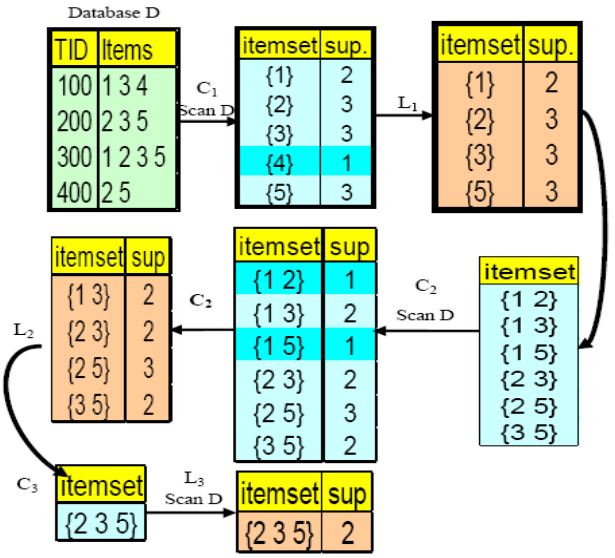

Fig. 2: A diagram showing scan database for frequent item sets (cse.ohio-state.edu, 010)

3. Priori algorithm

1. C_1 = Itemsets of size one in I;
2. Determine all large itemsets of size 1, L_1;
3. i = 1;
4. Repeat
5. i = i + 1;
6. C_i = Apriori-Gen(L_{i-1});
7. Count C_i to determine L_i;
8. until no more large itemsets found;

C. Clustering

Cluster analysis is the task of discovering groups and structure in the data that are in some way of "similar", without using known structure in the data. The goal is to partition a set of records into groups such that records within a group are similar to each other and records that belong to two different groups are dissimilar. Similarity between records is measured computationally by a distance function. (Ramakrishnan R et al. ,2003)

D. K – Means Clustering

K-Means methodology is a commonly used clustering technique. In this analysis the user starts with a collection of samples and attempts to group them into 'k' Number of Clusters based on certain specific distance measurements. (camo.com/resources/clustering.html, 2010).The K- means algorithm assigns each point to the cluster (also called centroir) is nearest. The center is the average of all the points in the points in the cluster- that is its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

Table 1: A Created Relational Database Table

Sid	Name	Age	GPA
1	John Barile	25	3.02
2	Mike Anikpo	23	2.84
3	Uche Madu	26	1.89
4	Nduka Irabo	29	3.75
5	Ebimene Oyeins	22	1.09

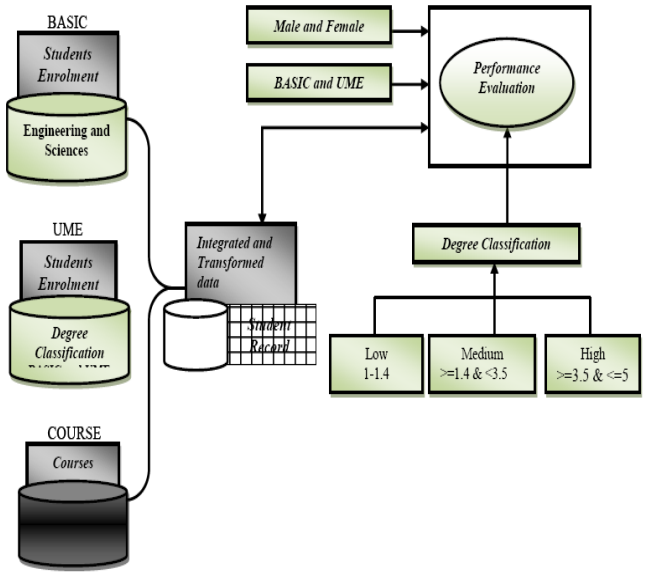

Fig. 3: Data Flow diagram showing Students Enrolment for BASIC and UME

Group Analysis

Class	Male		Female	
	No.	Percentage	No.	Percentage
High				
Middle				
Low				
Very Low				

Cluster Analysis

Male

Cluster Point	No	%

Female

Cluster Point	No	%

Fig. 4: Diagram showing Evaluation by Sex

Group Analysis

Class	Male		Female	
	No.	Percentage	No.	Percentage
High				
Middle				
Low				
Very Low				

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Cluster Analysis

BASIC

Cluster Point	No	%

UME

Cluster Point	No	%

Fig. 5: Diagram showing Evaluation by Admission

VI. OUTPUT SPECIFICATION

The figures 6-7 below show the results after evaluation of Students Performance using Data Mining Over a given data space.

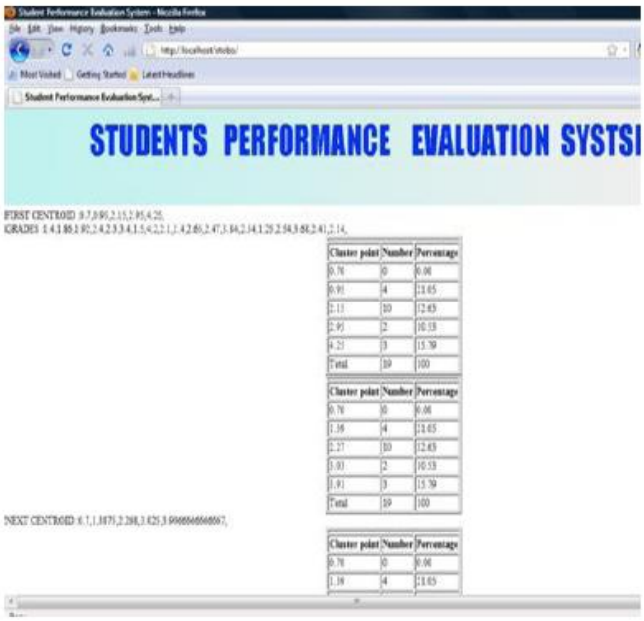

Fig. 6: Screenshot showing Cluster points, Number and Percentage of Student after Evaluation

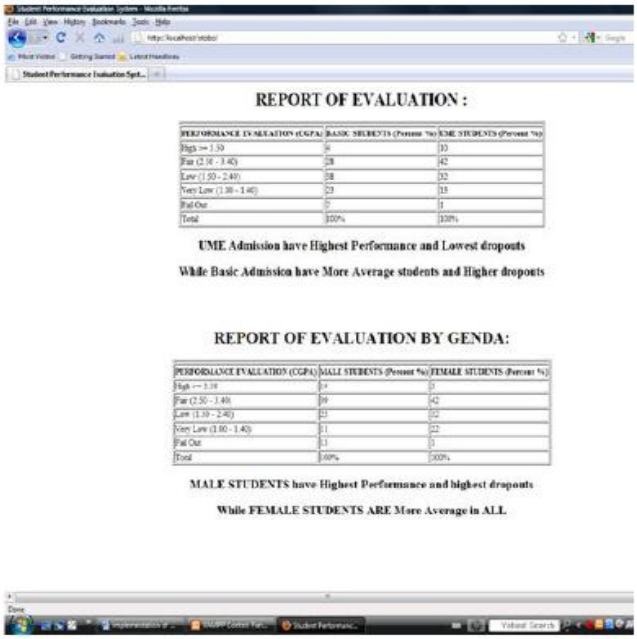

Fig. 7 Screenshot showing Report generated after Evaluation by GENDA

VII. CONCLUSION

This study introduced data mining technique 'K-means algorithm' to compare student performance. The data mining technique was applied in obtaining the required result from a huge amount of data stored in a database. The system currently was able to check and compare the performance of the students, the performance checked is based on student that where admitted by admission (BASIC and UME), by sex (male and female) and by degree ('High', 'middle', and 'low' performance) during their years of studies. By knowledge of their performance will enable the University to focus mostly in the area of the admission process via BASIC and UME.

RECOMMENDATION

We recommend the result of this research paper to the university of Port-Harcourt for further test in more data from the entire university and using the result to make informed decisions on the ration of admission between Basic and UME modes of intake. The research paper work can also form the fulcrum of other related research work in data mining, clustering and evaluation system development by other researchers. Students who are doing other research work in performance evaluation in data mining can find this research paper useful.

We also we want to recommend this to other universities who may want to use this as a prototype to achieve the in achieving same objective for good decision making based on their on admission processes in checking student performance in each course of study after leaving schools.

REFERENCES

1. Karoline Schönbrunn, Andreas Hilbert (2006), Data Mining in Higher Education.pg 489-496
2. Michael V. Mannino (2004), Database Design, Application Development, and Administration
3. Berzal, Fernando, Cubero, Juan-Carlos, Marin Nicolas, Serrano, Jose-Maria (2001). TBAR: An Efficient Method for Association Rule Mining in Relational Databases. Knowledge Engineering 37: 47-64.
4. Comaford, Christine. (1997). Unearthing Data Mining Methods, Myths. PC Week 14, no. 1 (January 6): 65.
5. Goebel, Michael and Gruenwald, Le. (1999), A Survey of Data Mining and Knowledge Discovery Tools. SIGKDD Explorations, ACM SIDKDD 1, no. 1 (June): 20-33.
6. Ramakrishna R. and Johannes G. (2003) Database Management System. McGraw-Hill
7. Ansari E., G.H. Dastghaibifard, M. Keshtkaran, H.Kaabi (2008).Distributed Frequent Itemset Mining using Trie Data Structure.