International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 06
May 2017

Monitoring Students' Academic Performance Using Clustering Algorithm Paradigm

Nwosu P. C[1].

Department of Computer Science
University of Port Harcourt

F[2]. E. Onuodu
Department of Computer Science
University of Port Harcourt, Nigeria

Abstract: *The capability of monitoring students' academic performance is a very big challenge to the academic community, especially institutions of higher education. A research work implore with K-means clustering algorithm method to monitor students' academic performance. The developed system was used to observe the progression of students' academic performance in the Department of Computer Science, University of Port Harcourt. The system was implemented using JAVA programming language and the methodology applied is Structured Systems Analysis and Design Methodology (SSADM) approach and MySQL as database. The result of this work will be useful in academic environment as performance monitoring and evaluation system tools. It will also enhance the decision making process of academic planners as the students' academic performance can be monitored session by session.*

1.0 Introduction

Grade Point Average (GPA) is a generally utilized as a medium for checking academic performance in tertiary institutions. Nigerian higher institutions indicate the lowest point that students are expected to maintain so as to progress in studying any given course. Most universities set the grade point of 1.5 as the minimum point . However, most advance degree programs such as Post Graduate Diploma (PGD) or M.Sc. graduate program may require a GPA from 3.0 and above as an assessment of good academic performance. Consequently, grade point average continues to be the widely accepted medium utilized by school administrators to assess students' academic progression. (Oyelade et al., 2010)

During the students' academic year, a number of issues could deter them from achieving or keeping up to a good academic record which eventually determines their grade point average. Academic administrators can focus on these issues and then provide means to increase and improve on the impact of academic activities and improve their academic performance by way of monitoring the progression of their performance. Probably, essential attributes can be discovered from students' academic performance using data mining tools like the clustering algorithm. The discovered attributes can then be utilized to determine the eventual overall performance of the students in order to improve or maintain good academic record of students.

Clustering algorithm assigns an array of objects into groups known as clusters such that the items that are in one cluster have more resemblance to one another than the ones in another cluster.

It is a major tool applied in data mining as well as a general method used in analyzing statistical data. Clustering is a very essential application area used in data mining. It is also applied in so many other domains of knowledge such as information retrieval, pattern recognition, image analysis, machine learning, bioinformatics etc. It provides a means of extracting of important information from extremely large amount of data, it is usually carried out in order to bring out unknown pattern and any existing interaction among objects valuable when taking decision on issues bothering around the data. (Shovon et al., 2012)

2.0 Related Works

Clustering is a task that arranges a set of sample items into groups known as clusters such that the items found in a particular group are most related in any way to one another than to those found in other groups.

A simple analytical explanation of clustering is given as follows: assume $X \in R^{m \times n}$ is a set of data items defining a set of m points X_i in R^n. The objective is to divide x into k clusters c_k so that all the items belonging to a particular cluster are most related than items in another cluster. (Fung, 2001)

where:

- $X \in R^{m \times n}$ represents a set of data objects defining a set of m points.
- C_k represents the k^{th} cluster.

- *K* represents the total number of clusters.
- C_k^j represents the k[th] cluster center at the iteration j.

Types of Clustering Algorithms

Hierarchical Clustering

The hierarchical algorithms generate an orderly breakdown of the items (Andritsos, 2002). This type of algorithm is of two types; the agglomerative and the divisive algorithms.

(a) The Agglomerative algorithm: This type of algorithm begins with every item as a independent cluster, then it merges the clusters in succession in line with a distance measure. Grouping can be discontinued whenever a desired result is achieved or when the entire items appear in one cluster. This type of technique normally follow a particular order starting from down then moves upward, therefore it is also known as the bottom-up algorithm.

(a) The Divisive algorithm: The divisive method just the reverse of the agglomerative method. It usually begin a single cluster that contains all the objects, then repeatedly divide to form smaller clusters, it continues tile every item appear in a group or when an expected result if reached. It is like the divide-and-conquer algorithm method, it splits the data items into separate cluster at each point, and continues till the entire items appear in one cluster.

Partitional Clustering

Partitional clustering divides a set of data items into groups, with each of these groups correspond to a cluster. The grouping is on an objective function. which is minimizing square error function. (Elavarasi et al, 2011)

Partitional algorithms seeks to locally enhance a particular condition. The values of the similarity initially computed, they order the results, and pick the one that optimizes the criterion. Hence, the majority of them could be considered as greedy-like algorithms. (Andritsos, 2002)

(a) K-Means Algorithm

The K-means clustering is generally used because it is simple to apply and its has very good performance many areas. It is one of the first ten powerful clustering algorithms applied in data mining.

This type of algorithm requires the reduction of the sum of the squared-error function which is actually easy to carry out. Nevertheless, some disadvantages are attached; the number of clusters in a given dataset has to be previously known, the final outcome of the algorithm is mainly dependent on the initial samples, it is sensitive to irregularities or anomalies, etc. The K-means algorithm functions well for globular pattern clusters, as well as similar size and massive clusters.

A simple K-means algorithm is given as follows (Mahdi et al., 2010):

Step1: Choose *k* items as preliminary centres.
Step 2: Allocate each data object to the closest centre.
Step 3: Recalculate the centres of each cluster.
Step 4: Repeat steps 2 and 3 until centres remain constant.

(b) **The K-modes Algorithm**

This algorithm is an enhancement of the K-means clustering algorithm, the difference is that the mode of the items is used instead of the mean. This technique was adopted in order to take care of categorical elements. The mean of the cluster as applied in the k-mean method are replaced with the mode. During the clustering process, minimizing the clustering objective function ensured using a density-based approach to update modes in the. The modes are the attribute values with high frequency. (Huang, 1998)

(c) **The K-prototypes Algorithm:** This algorithm was proposed to handle clustering of large datasets involving mixed attributes. It is essentially valuable as real world objects found in databases are usually of mixed-type; they often contain both numerical and categorical data. (Haung,1998)

Numeric and categorical values are usually well handled by dissimilarity measure. Take for example, if dissimilarity measure is given by *x*, the dissimilarity measure on numeric attributes defined by the squared euclidean distance is s_n while the dissimilarity measure on categorical attributes defined as the number of mismatches of categories between two objects is s_c. Then, the dissimilarity metric between the two items will be given by $\boldsymbol{x = s_n + \gamma s_c}$, where γ represents the weight for balancing the two clustering process. A major issue in using the algorithm is in selecting an appropriate weight; thus, average standard deviation of numeric elements should be applied when selecting the weight.

$$x = s_n + \gamma s_c \quad (1)$$

where x =dismilarity measure on numeric attributes.
s_n = dismilarity measure on categorical attributes.
γ = weight for process balancing.

(d) Partitioning Around Medoids (PAM) Algorithm

This algorithm is an enhancement to K-means clustering algorithm proposed to efficiently control irregularities. It assigns each cluster by its medoid instead of cluster centers. And it is normally applied on data when there is difficulty in determining its mean or center. The medoid refers to item that most centrally located in a cluster, as a result, medoids not affected by extreme values. The PAM clustering method initially selects *k* medoids and then seek to position all other items whose medoid is nearer in a cluster, substituting medoids with non-medoids thereby improving the quality of the outcome. This quality can be assessed by applying the squared-error function between its medoid and the items in a cluster. The computational complexity of PAM is defined by $O\{I\ k(n-k)^2\}$, here, I represents the number of repetitions, thus the computational cost is extremely increased when the values of n and k are high. (Androtis, 2002)

(e) Spectral Clustering Algorithm

The algorithm is mostly applied in Statistics, Engineering, Applied Mathematics but in recent times it has been as well adopted for various purpose in Computer Science problems. It utilizes eigenvalues and eigenvectors. It actually depends on a set of eigen values of a relationship matrix. It form clusters by splitting the data items using this relationship matrix (Elavarasi et al, 2011). This type of algorithm is normally associated with three main phases which include:

i. **Preprocessing:** This phase handles the structuring of the relationship matrix.
ii. **Spectral Mapping:** This phase handles the structuring of eigen vectors for the relationship matrix.
iii. **Post Processing:** This phase handles of clustering the data items.

Nevertheless, a major problem with this method is that it presents a great calculation difficulty. When larger dataset is involved, the method requires O(n3); where n represents the number of data items. Some examples of spectral algorithm include the Linkage algorithm, the Kannan, Vempala and Vetta Algorithm (KVV), The Ng, Jordan and Weiss (NJW) algorithm, the Shi and Malik (SM) algorithm.

Similarity/Distance Measures

Grouping the samples is based on similarity or distance measure. This deals with the closeness of the objects and can be calculated using several methods. Some are euclidean in nature, that is to say it can be obtained with a measuring instrument like ruler. and there are other distances based on similarity. There are other distance measures that can be utilized but the squared-euclidean distance is widely used. (Fung, 2001)

(a) Euclidean distance

This refers to is the normal geometric distance between two points in a metric space. Assuming there are objects, the euclidean distance is the mathematical difference between the values obtained by taking the measurement of the points with an ordinary ruler. The euclidean distance is a appropriate measurement parameter (Hoon et al, 2002), as it satisfies the triangle inequality as:

$$d = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \quad (2)$$

where x_i and y_i are initial distance of two points

(b) Squared Euclidean Distance

The simplest way as well as widely acknowledged method of calculating the distances between objects in a multi-dimensional space is to compute the euclidean distances. But the squared euclidean metrics is normally utilized so that larger weights are gradually positioned on objects that are farther away. This is similar to the ordinary euclidean measure only that that the distance is calculated without taking the square root of the values, which makes it faster for clustering. The euclidean metrics are usually calculated using unprocessed data.

(c) Minkowski Distance

This is a metric used to determine the similarity or dissimilarity that exist among objects. The Minkowski metric defines the distance between object in a vector space where the norms are defined.

(d) Mahalanobis Distance

This was introduced by P. C. Mahalanobis, the algorithm is a distance metric between two point. It

removes the distance distortion caused by linear combinations of attributes.

(e) Cosine Distance

Cosine metric is a measure of resemblance that exist between two arrays of numbers by obtaining the vectors of the points and computing the cosine of the angle between the points using their dot products. It can be mathematically defined by:

$$d_{cos}(X,Y) = \frac{X.Y}{|X||Y|} \quad (3)$$

Where x and y are distance, x.y is the dot product.

Clustering Process

Clustering definition implies that clustering involves difference processes, (Jain et al, 1988). These processes can be divided into:

i. **Data Collection**: This is the first task to be carried out and it involves the cautious mining of relevant data items from a primary source of data.
ii. **Initial Screening**: At this stage, the data is fine-tuned after obtaining if from the primary source.
iii. **Representation**: This process involves appropriate organization of the available data so as to make them fit for use in a clustering algorithm. Similarity measure is selected and then the data attributes and measurement are checked.
iv. **Clustering Tendency**: This process ensures that the underlying data has propensity to form clusters naturally. But often times, the step is overlooked, particularly when huge datasets are involved.
v. **Clustering Strategy**: This step involves cautious selection of clustering method to be applied and the preliminary boundaries.
vi. **Validation**: According to estimation, this is one of the stages typically not well represented. It is usually centered on physical assessment as well as visible procedures. Nevertheless, with increasing amount of data and their measurement there is no way of result evaluation with predetermined concept or clustering.
vii. **Interpretation**: It involves the merging of other analysis together with the clustering results, so as to make deduction and recommend any additional analysis that may be required. Interpretation offers anticipated result for the previous stages. It usually begins with data gathering and assessment.

3.0 Materials and Method

3.1 Analysis and Design

Educational institutions work to improve on the students' academic performance as well as the overall other extra-curricular activities but with a large student body it is always a difficult challenge to evaluate the impacts of curriculum changes.

The manual ways of evaluating or monitoring students' performance is by counting and separating student's performance scores or using Microsoft Excel to sort before counting but this process is time-consuming and tedious because of the amount of data involved.

The ever increasing amount of data usually constitutes a huge task which makes it needful to provide tools for analyzing these data in order to discover similarities in them. The process of data mining was developed as a discipline that supply these tools for data analysis, ascertaining unknown facts as well as taking independent decision for various application areas.

The proposed system uses K-means clustering algorithm to monitor students' academic performance. To effectively apply a data mining tool, database must be involved, so the first thing is to obtain the data for mining and then create a database using these data. In this research work, the database was constructed using the information acquired from the final year students of Computer Science department results.

The K-means algorithm divides "n" samples into k clusters where every sample is part of the cluster with the nearest mean. This algorithm tries to minimize the squared error function.

The Object-Oriented Analysis and Design Method (OOADM) was adopted in the design to monitor the academic performance of the students.

The software requirements for this system are:

- MS-Windows Operating System (at least Windows XP)
- Java
- SPSS 20.0

- MySql Server
- Netbean

K-Means Clustering Algorithm

The algorithm for K-means clustering method is outlined as follows:

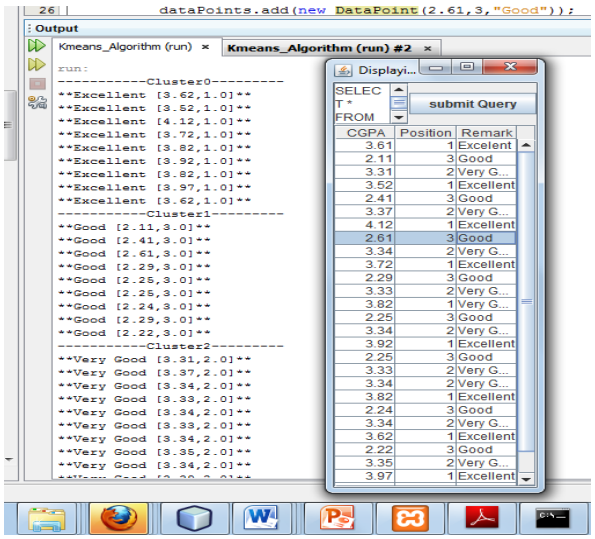
1. Choose *k* point as the preliminary centre.
2. Build *k* clusters by allocating all items to the nearest centre.
3. Re-calculate the centre of every cluster.
4. Until the centres remain constant.

4.0 Result and Discussion

The outcome of applying K-means algorithm on student dataset collected from 2008 academic session of Computer Sciences Department University of Port-Harcourt was presented. The dataset contains 100 instances which were randomly selected. Every instance contains three attributes which include CGPA, Position and Remarks; the attributes are numerical.

The process randomly splits the dataset into clusters and then creates the categories depending on the average of the dataset content. The relevance of the objects to the created cluster is envisioned each time a new object is added using the categories of these dataset which comprises of three categories which are the following:

i. Good Students
ii. Very Good Students
iii. Excellent Students



A **Sample of the Result**

Finally, the proposed system was able to cluster all the data in the dataset irrespective of the position of the particular data in the database and the results are displayed.

5.0 Conclusion

Evaluating academic performance of students' posses a big challenge to academic institutions, especially institutions of higher learning. Clustering can be efficiently utilized in academic institutions for evaluating the academic performance of students. The research work introduced the K-means clustering algorithm for assessing students' academic performance in institution of higher learning to evaluate the relevance of new dataset introduced into an existing ones. The whole dataset utilized in the experiment are of numerical types involving scores in former semester including examinations, assignments, term papers and class projects scores. The analysis result on the datasets shows that the K-means algorithm performs well.

The algorithm was applied to obtain the grouping of the newly introduced items in the dataset at a little computing charge. This study proposed an efficient application that integrates fields of knowledge and analyzes methods to foretell the outcomes of future challenges involving students' academic performance .

6.0 Recommendation

For future work, the method will be enhanced to have an efficient system with additional features so as to obtain more reliable and correct results valuable to academic administrators for improving students learning outcomes. Consequently, the k-means clustering algorithm provides an excellent scale for observing the performance students' of students as they advance in their studies. Decision making by academic administrators can be improved academic record can be observed sessions by session which in turn helps to improve or maintain future academic performance in subsequent academic sessions.

References

1. Andritsos P. (2002) "Data Clustering Techniques". Qualifying Oral Examination Paper, Department of

Computer Science, University of Toronto, Canada. Vol.3(8), 146-149.

2. Elavarasi S. A, Akilandeswari J. and Sathiyabhama B. (2011), "A Survey on Partition Clustering Algorithms". An online International Journal of Enterprise Computing and Business Systems Vol 1(1). Retrieved on 04 October, 2012 from www.ijecbs.com.

3. Fung G. (2001) "A Comprehensive Overview of Basic Clustering Algorithms". A Technical Report, University of Wisconsin, Madison. Vol. 31 (3), 24-33

4. Hoon M. J., Imoto S. and Miyano S. (2010) "The C Clustering Library for DNA Microarray Data" Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo, Japan. Vol. 31 (3), 164-223

5. Huang Z. (1997) "A Fast Clustering Algorithm to cluster Very Large Categorical Datasets in Data Mining", In Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Vol. 3 (3), 64-83

6. Huang Z. (1998) "Extensions to the K-means algorithm for clustering Large Data Sets with Categorical Value", Data Mining and Knowledge Discovery September 1998, Volume 2, (3), 283–304, Kluwer Academic Publishers, The Nertherlands.

7. Huang Z. (2003) "Clustering Large Data Sets with Mixed Numeric and Categorical Values". In Proceedings of The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 105-108, Allied Publishers PVT Ltd., Mayapuri, New Delhi.

8. Jain A., Murty M., and Flynn P. (1999) "Data Clustering: A Review", ACM Computing Surveys (CSUR), New York, NY, USA. Vol. 31 (3), 264-323.

9. Oyelade O. J., Oladipupo O. O. and Obagbuwa I. C. (2010) "Application of K-Means Clustering Algorithm for Prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, Vol. 7, (1), 292-295.

10. Shovon I. H. & Haque M. (2012), "An Approach of Improving Student's Academic Performance by using K-means Clustering Algorithm and Decision Tree", International Journal of Advanced Computer Science and Applications, Vol.3(8), 146-149.