HYBRID DATA MINING MODEL FOR KNOWLEDGE DISCOVERY ON STUDENTS ACADEMIC PERFORMANCE

Chinenye C. Opara[a], U.F. Eze[b], Chukwuemeka P. Oleji[c]

[a,b]*Department of Information Technology, School of Computing and Infromation Technology, Federal University of Technology Owerri,*
[c]*Department of Computer Science, School of Computing and Infromation Technology, Federal University of Technology Owerri*

[a]*mirbelchinenyec@gmail.com*

Abstract

Over the years, large record of student data exists in Institutions of higher learning because students graduate from these institutions yearly. This has necessitated the need to explore those data to discover some patterns and relationships existing in them and as well make strategic decisions for a better education system. This paper focused on developing an enhanced Hybrid data mining model to mine students' progress and performance for knowledge discovery and for decision making purposes. This was implemented using Clustering Algorithms (k-means and k-representative) as a tool in data mining to group students data associated to its grades. Object-Oriented Analysis and Design Methodology, NetBeans Integrated Development Environment (IDE) and Matlab editor were all adopted for the analysis and design of the proposed system respectively. The proposed hybrid model efficiently clustered mixed dataset of past graduated students record which was used in this work and the percentage distribution of knowledge discovered indicated a poor performance level of students. The result also shows that the hybrid clustering algorithm improved K-means clustering algorithm for optimal solution and efficient clustering of mixed data sets with 99% performance and clustering error of 0.0025. Therefore, the result of this work will enable Academic planners to efficiently monitor students performance based on the performance level of students and also make decisions for a better education system.

Keywords: Academic Performance, Clustering Algorithm, Data mining, Hybrid Data Mining Model;

1. Introduction

Academic performance is seen as the result or achievement of once educational progress, and can also be attributed to the extent to which Institutions and students have achieved a successful educational goal. Recently, there has been a noticeable decline in the quality of graduates of some Nigerian Universities which enforces cost on the society and the country at large because these students are seen to be the key players in the affairs of the country and all its economic sectors. Therefore, Students' academic performance should be monitored closely but the ability to monitor the progress and as well make strategic decisions for a better education system is a critical issue in most Institutions of higher learning. Although, a lot of factors such as family problems, finances, social life, poor communication, etc. could determine the students' academic performance, but Grade Point Average (GPA) is a common indicator to determine students' performance in Institutions of higher learning and that is the raw data to this study. Data mining also known as knowledge discovery on database is a tool used to discover hidden patterns and relationship existing in large data set, was implemented in this study.

The main function of data mining is to apply various methods and algorithms to discover and extract novel information on stored data. (Shirwaikar & Rajadhyax, 2012). As a result of this, a lot of research interest is

drawn on the implementation of data mining in the area of academics and educational system as a whole. A lot of data mining techniques such as clustering, Decision tree, Neural networks, Naïve Bayes, k-nearest, etc. have become the center of interest in research. Clustering analysis as an important tool in data mining was used to developing the hybrid model for knowledge discovery on students' academic performance. K-means clustering algorithm is known for its efficiency in clustering large data sets but it is limited to clustering only numerical data sets, therefore making it less relevant in clustering real word datasets. In this research, we implemented k-representative as an extension of k-means to improve the clustering accuracy and as well, cluster categorical or mixed datasets by using a simple matching dissimilarity measure and frequency-based method. These two clustering algorithms were grouped to develop the Hybrid data mining model for knowledge discovery on large volume of students record datasets, in order to make useful and constructive decisions by the academic planners of the Universities in Nigeria.

1.1. Data Mining Concepts

For Data Mining (sometimes called knowledge discovery on data) is a process in which data is analyzed and summarized to generate useful information that can be used for different purposes such as monitoring process, weather forecasting, anomaly detection, fraud detection, pattern recognition and so on. (Shruthi & Chaitra, 2016). Data mining technique is one of the major analytical tools for analyzing data from different perspectives by identifying the relations existing among them.

1.2. Concepts of Clustering

Clustering is a data mining technique used to group data objects similar to one another within the same cluster and dissimilar to the objects in another cluster. Clustering is applied in different areas including pattern recognition, data analysis, image processing, and market research. Data clustering is alternatively referred to an unsupervised learning and statistical data analysis which is also an important analysis of human activity. Clustering is a descriptive task that seeks to identify homogenous group objects based on the values of their attributes.(Govindasamy, 2018). There are different types of clustering algorithms but the two most widely used clustering algorithm is K-means clustering algorithm which uses the centroid model and hierarchical clustering algorithm which uses connectivity model.

The K-means clustering technique is an iterative algorithm where items are moved among sets of clusters until the desired are related. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously. The K-Means clustering technique is used to classify data in a crisp sense. Define a family of sets { i , 1,2,3, , } Ai = · C as a partition of X,

where the following set-theoretic forms apply to those partitions.(Patel & Yadav, 2015).

2. Literature Review

In the research by (Oyelede et al.,2010), students' result was analyzed using clustering and standard statistical algorithms. k-mean clustering algorithm was used to cluster students' result data. The model was combined with the deterministic model which was a good benchmark to monitor the progression of academic performance but is limited to only numerical data. On the other hand, in the work of (Kumar and Baradwaj, 2011), the authors used classification to evaluate student's performance, the decision tree method was also used. Based on their result, they extracted knowledge that describes students' performance in the end of semester examination. In another work of (Mishra and Sangeeta 2014) classification techniques were used to build performance prediction model based on students' social integration, academic integration, and various emotional skills. Two algorithms J48 (Implementation of C4.5) and Random Tree was applied to predict their

performance. Also, (James and Hary 2014) demonstrated the use of cluster analysis in students' results and using statistical algorithms to segregate their marks based on their performance. They also integrated k-means clustering algorithm and deterministic model to analyze the student's results and their performance was based on a test score by the students. (Nikhil and Rudresh 2015) on the other hand, predicted Students' Performance using Frequent Item Set Mining, Clustering & Classification. They focused on classifying students into different categories such as good, average, and poor depending on their marks scored by them. They also integrated decision tree which greatly improved the prediction on the performance of the students. (Rana and Garg, 2016) also evaluated Student's Performance of an Institute Algorithms using K-means and hierarchical clustering algorithm. A comparison was made between two unsupervised algorithms using WEKA Tool as an open-source tool. Furthermore, (Pat et al. 2016) focused on analyzing student performance using fuzzy logic expert system and fuzzy k-means expert system. Fuzzy k-means was used to predict the students' performance while fuzzy logic was used to convert the crisp data into fuzzy sets and computes the total marks of a student. In the work of (Govidasamy, 2018), the author evaluated student's academic performance using k-Means, k-Medoids, Fuzzy C Means (FCM) and Expectation Maximization (EM). The performance of the clustering algorithm was compared based on the factors: Purity, normalized mutual information (NMI) and time taken to form clusters. The results show that FCM and EM algorithm performs well compared with the other two clustering algorithms. Moreover, (Nagesh et al, 2018) analyzed student data and predicted the percentage of students whose academic performance is poor, average and good by making use of k-means clustering algorithm. The 50% of the data clustered as good shown in green wish was illustrated in their output graph. (Aggarwal and Sharma, 2019) analysed students performance based on the actual result of university examination using k-means clustering algorithm. The elbow method was also used to choose the appropriate number of clusters. Their results showed that more female in cluster 2 and 3 had good performance and better academic performance as compared to male students.

3. Analysis and Design

The Hybrid data mining model is analyzed using the Data flow model shown in figure 3.1

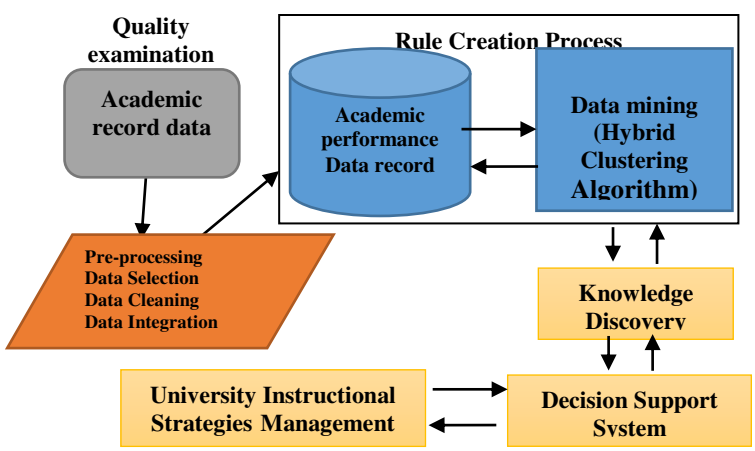

Fig 1. Flow Diagram of the Proposed Hybrid Model

The flow diagram shows a well structural flow of different functional modules and phases of the system as follows;

- Academic record data: Relevant data is been retrieved and analyzed from the students academic record dataset.
- Data Preprocessing Stage: The second phase of knowledge extraction has to do with the combination of both data integration and data cleaning. Data cleaning requires removal of irrelevant and unwanted data, while data integration is phase has to do with the integration of multiple data sources, often heterogeneous, into a common source appropriate for the mining procedure.
- Data Mining: This is an important stage where useful patterns are discovered from the pre-processed students academic record data using sophisticated techniques. The techniques applied here is clustering algorithm where k-means and k-representative clustering algorithm were integrated to cluster mixed dataset and knowledge was extracted based on the data mining technique to be used.
- Interpretation and Evaluation: In this phase, related patterns which represent knowledge is been identified. This is done based on the given measures below: Decision Support System (DSS), Knowledge Discovery on Database (KDD), Instructional Database and Academic Performance

3.1. Data Processing Method

This was achieved through primary data, observation, and interview, after which critical analysis of the students grading system was evaluated. The raw data for this study was retrieved from the statistics and record unit of Federal University of Technology Owerri as one of the Universities under the Nigerian University Commission (NUC). The data comprises of students' gender, cumulative grade point average (CGPA) and the performance. The experimental data specification is illustrated in table 3.1.

Table 1. Data specification table

Domain	CGPA	Gender
1st Class	4.5 - 5.0	Male
2nd Class Upper	3.5 – 4.0	Female
2nd Class Lower	2.4 – 3.4	

The datasets used in this work were stored in MATLAB editor for easy mathematical computation of the proposed algorithm. The database table contains three columns, field name, type and description. Field name consists of Position and students Cumulative Grade Point Averages (CGPA), the data type for remark is text. The data type for students' CGPA is float. The float type for gender is numeric (1 = male, 2= female). The objective is to enhance technical design specifications for a database, which can adapt to future requirements and expansion

3.2. Proposed Hybrid Model Specification

References In the proposed model, we integrated K-means and K-representative clustering algorithm as an efficient clustering technique for the hybrid data mining model. K-means clustering algorithm is noted for clustering large data sets but limited to clustering only numerical data. K-representative algorithm was implemented to enhance the clustering accuracy of K-means by using a simple matching dissimilarity measure and a frequency-based method to update its means to mode. These two algorithms were integrated to develop the hybrid data mining model which was used to evaluate the student's performance and as well used the derived knowledge for decision making. The model is described as follows

A. K-means clustering Algorithm

K-means clustering Algorithm is aimed at clustering the best center called centroid in a given data set. The idea is to initialize k centroid, one for each cluster, then the distance of all data elements are calculated using Euclidean distance measure such that data sets closer to the centroid are clustered together. This step is repeated until no more changes occur in the cluster. The procedure for k-means clustering algorithm is as follows:

INPUT: Number of k initial clusters
Data objects D= {d1, d2…dn}
OUTPUT: A set of K clusters
Steps:

1. Select K initial centroid from D
2. Repeat,
3. Calculate the distance between each data objects Di(1<= i<=n) to the closest cluster k
4. For each cluster j(1<=j <=k), recalculate the cluster centroid
5. Until the centroid do not change

i. Mathematical Model for K-means Clustering Algorithm

$$d_2(X,Y)\sum_{j=1}^{p}\left(x_j - y_j\right) \qquad (1)$$

The above model describes the distance between two datasets X and Y, while ∑ denotes the summation of clusters for the numerical datasets.

B. K-representative Clustering Algorithm

K-representative clustering Algorithm was used in this study to group n objects based on attributes into k number of groups, each attribute belongs to the cluster with nearest mean, where k is a positive integer. The grouping is done by minimizing the sum of squares of distance between data and corresponding cluster centroid. The steps for k-representative clustering algorithm is a follows:

INPUT: Number k initial centroid
Data objects D= {Xi, Yi…n}
Steps:

1. Randomly select k initial centroid
2. Calculate the k-representative for each cluster Ci
3. For each object, calculate dissimilarity d(Xi=Yi=1), reassign Xi to cluster C Update both Xi and Yi
4. Repeat step 3
5. Until no object has changed within the cluster

ii. Mathematical Model for K-representative clustering Algorithm

$$\sum_{j=p+1}^{m} cf\delta\left(x_j, y_j\right) \qquad (2)$$

The model above shows the distance of two data objects X and Y of two data objects X and Y using a frequency base measure for the categorical attributes associated to the datasets.

3.3. Mathematical Model of the Hybrid Algorithm

A. Input Model

The design of the system requires a complete understanding and processing of the domain problem. In this step, the input data is retrieved, modelled and computed using the mathematical model below:

$$d_2\left(X,Y\right) = \sum_{j=1}^{p}\left(x_j - y_j\right)^2 + \gamma \sum_{j=p+1}^{m} cf\delta\left(x_j, y_j\right) \qquad (3)$$

The above input model shows the dissimilarity between two data sets X and Y which is seamlessly summed based on both numerical and categorical objects. The first term denotes the numerical datasets while the second term denotes the simple matching dissimilarity measure on categorical attributes. X and Y represents data objects, d2 shows the distance between two objects X and Y, $\sum_{i=1}^{p}(x-y)$ represents the sum of all data objects while 2 represents gender, cf is the cumulative frequency measure for categorical datasets and γ is the weighted measure balance between numerical and categorical objects. Hence, in the model, the numerical and categorical objects are hybridized to develop the hybrid data mining model for knowledge discovery based on the input data

3.4. Design of the Hybrid Algorithm

B. Output Design

NetBeans Integrated Development Environment 8.0 (IDE) was used to develop and implement the system which shows various distinct clusters of the attribute when executed. The number of clusters is determined by the value of K as shown in figure 3.2.

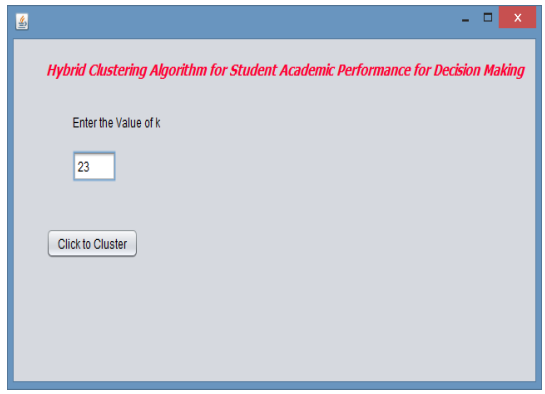

Figure 2. Output Design on NetBean IDE

The output result displays the clustered Cumulative Grade Point Average (CGPA) of students which consists number of clusters indicating students CGPA and remark (first class, second class upper, second class lower and third class). The result is shown in figure 3.

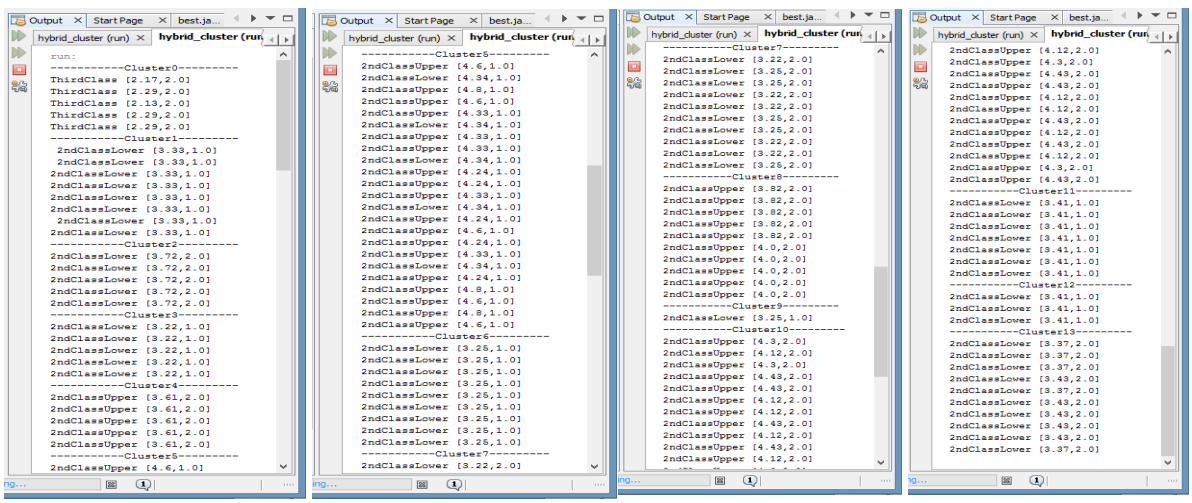

Figure 3. Output Design of Clustered results

4. Results and Discussion

The results showed that instances of different classes were separated by clusters successfully as shown in output results in figure 3. Moreover, the number of clusters were adjusted through removing unnecessary clusters during running of the algorithm. We used mix classification matrix in table 2, otherwise called confusion matrix to measure the accuracy of the clustering algorithm from the total number of 2000 students record datasets. The confusion matrix outlined the number of instances of a particular class or attribute such that the number of clusters for First class is 174, Second class upper is 531, Second class lower is 714 and Third class is 581. This represents a complete recovery of the number of instances of data in a cluster from the

2000 students record data.

Table 2. Mix Classification Table

Merged Clusters	Merged Clusters	Merged Clusters	Merged Cluster	Merged Clusters	Merged Clusters
First Class	174	0	0	0	174
Second class upper	0	531	0	0	531
Second class lower	0	0	714	0	714
Third class	5	0	0	576	581
					200

The graph and the corresponding table denote the percentage distribution of knowledge discovery on mining students performance using the proposed model. From table 2, we observed that the First Class is 8.7%, Second Class Upper is 25.55%, Second Class Lower is 35.7% and Third Class is 29.05%. The result in Figure 4, shows a level of poor performance on the students.

Table 3. Percentage distribution of knowledge table

Remark	Total Clusters	Academic Performance (%)
First Class	174	0
Second class upper	0	531
Second class lower	0	0
Third class	5	0

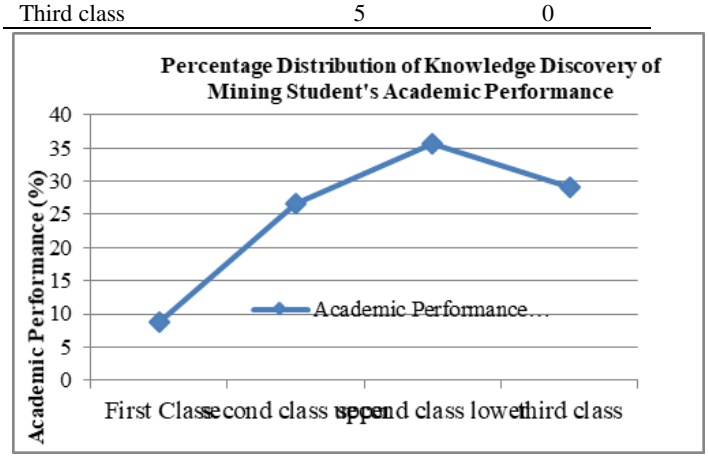

Figure 4. Percentage Distribution Graph

The results were tested using clustering accuracy measure as proposed by (Oyalade et al 2010) with the formula;

$$r = \frac{1}{n}\sum_{l=1}^{k} a_l,$$

where a; denotes the number of data objects which occur in both cluster and its corresponding class, n; gives the number of objects in the datasets.
Hence, the clustering error is defined as e = 1- r
r = (174 + 531 + 714 + 576) / 2000 = 09975 and e = 0.0025
From the above clustering error, it shows that the proposed hybrid clustering algorithm improved K-means clustering algorithm for optimal solution and efficient clustering of mixed data sets with clustering error of 0.0025.

5. Conclusion

In this paper, we were able to develop a hybrid model to mine students academic performance for decision making purposes using k-means and k-representative clustering algorithm. The proposed hybrid model provided efficient algorithm for clustering students' academic performance for knowledge discovery. And it improved K-means clustering algorithm for optimal solution and efficient clustering of mixed data sets with 99% performance and clustering error of 0.0025 based on the mix classification matrix table. From the result, the proposed hybrid clustering algorithm efficiently mine students' academic performance for constructive decision making strategies. Therefore, we recommended that educational management system should implement the results of this work in their education performance monitoring and assessments, which will help provide the universities with the status, possible problems and ability of the students. Lecturers and Academic advisers will equally teach and guide the students to achieve better performance.

References

Aggarwal, D., & Sharma, D., 2019. Application of Clustering for Student Result Analysis. 6, 50–53.

Govindasamy, K., 2018. Analysis of student academic performance using. 119(15), 309–323.

Jamesmanoharan, J., S. Hari Ganesh, M. Felciah, and A. K. Shafreenbanu (2014) "Discovering Students' Academic Performance Based on GPA Using K-Means Clustering Algorithm." In Computing and Communication Technologies (WCCCT), 2014 World Congress on 200-202. IEEE

Kumar, B., & Pal, S., 2011. Mining Educational Data to Analyze Students Performance. International Journal of Advanced Computer Science and Applications, 2(6). https://doi.org/10.14569/ijacsa.2011.020609

Mishra, T., Kumar, D., & Gupta, S, 2014. Mining students' data for prediction performance. International Conference on Advanced Computing and Communication Technologies, ACCT, February, 255–262. https://doi.org/10.1109/ACCT.2014.105

Nagesh, A. S., & Satyamurty, C. V. S, 2018. Application of clustering algorithm for analysis of Student Academic Performance. International Journal of Computer Sciences and Engineering, 6(1), 381–384. https://doi.org/10.26438/ijcse/v6i1.381384

Nikhil Rajahyax and Rudresh S, 2012. Data Mining on Educational Domain Journal Research of Computer Science. June 2012. 56-63

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C, 2010. Application of k Means Clustering algorithm for prediction of Students Academic Performance. 7, 292–295. http://arxiv.org/abs/1002.2425

Patel, J., & Yadav, R. S, 2015. Applications of Clustering Algorithms in Academic Performance Evaluation. OALib, 02(08), 1–14. https://doi.org/10.4236/oalib.1101623

Rana, S., & Garg, R, 2016. Evaluation of student's performance of an institute using clustering algorithms. International Journal of Applied Engineering Research, 11(5), 3605–3609.

Shirwaikar, R., & Rajadhyax, N, 2012. Analyzing Students Performance Using Frequent Item Set Mining, Clustering & Classification. International Journal of Management & Information Technology, 1(2), 31–41. https://doi.org/10.24297/ijmit.v1i2.1444

Shruthi, P., & Chaitra, B. P, 2016. Student Performance Prediction in Education Sector Using Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering, 6(3), 212–218.

Shiwani, R. and Roopali, G, 2016. Evaluation of Student's Performance of an Institute Using Clustering Algorithms. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 5 (2016) 3605-3609© Research India http://www.ripublication.com.