Computer and Information Science; Vol. 12, No. 2; 2019
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Unsupervised Characterization and Visualization of Students' Academic Performance Features

Udoinyang G. Inyang[1], Uduak A. Umoh[1], Ifeoma C. Nnaemeka[1] & Samuel A. Robinson[1]

[1] Department of Computer Science, University of Uyo, Uyo, Nigeria

Correspondence: Udoinyang G. Inyang, Department of Computer Science, University of Uyo, Uyo, Nigeria. E-mail: udoinyanginyang@uniuyo.edu.ng; udoiinyang@yahoo.com

Received: January 10, 2019 Accepted: February 27, 2019 Online Published: April 30, 2019

doi:10.5539/cis.v12n2p103 URL: https://doi.org/10.5539/cis.v12n2p103

Abstract

The large nature of students' dataset has made it difficult to find patterns associated with students' academic performance (AP) using conventional methods. This has increased the rate of drop-outs, graduands with weak class of degree (CoD) and students that spend more than the minimum stipulated duration of studies. It is necessary to determine students' AP using educational data mining (EDM) tools in order to know students who are likely to perform poorly at an early stage of their studies. This paper explores k-means and self-organizing map (SOM) in mining pieces of knowledge relating to the natural number of clusters in students' dataset and the association of the input features using selected demographic, pre-admission and first year performance. Matlab 2015a was the programming environment and the dataset consists of nine sets of computer science graduands. Cluster validity assessment with k-means discovered four (4) clusters with correlation metric yielding the highest mean silhouette value of 0.5912. SOM provided an hexagonal grid visual of feature component planes and scatter plots of each significant input attribute. The result shows that the significant attributes were highly correlated with each other except entry mode (EM), indicating that the impact of EM on CoD varies with students irrespective of mode of admission. Also, four distinct clusters were also discovered in the dataset by SOM —7.7% belonging to cluster 1 (first class), and 25% for cluster 2 (2nd class Upper) while Clusters 3 and 4 had 35% proportion each. This validates the results of k-means and further confirms the importance of early detection of students' AP and confirms the effectiveness of SOM as a cluster validity tool. As further work, the labels from SOM will be associated with records in the dataset for association rule mining, supervised learning and prediction of students' AP.

Keywords: students' academic performance, self organizing map, at-risk students, k-means, cluster analysis, cluster validity

1. Introduction

Education is one of the critical determinants of the quality of human capital and economic development. It is a major asset in both short term and long-term productivity, and advancement at both micro and macro levels of economic performance. Education enriches skills, abilities, intelligence quotient and qualifications of a country's work force and produces economic growth and high standard of living (Durkaya and Hüsnüoğlu, 2018). It enhances productivity, innovativeness, entrepreneurship and technological advances. It is one of the largest and visible sectors in the world, which attracts attention and huge resources to guarantee sustainable economic development and high standard of living (Vanthienen and Witte, 2017). The educational sector has advanced over the years, moving from manual techniques of data capture, collection, processing and dissemination, to methods driven by information technology, in the schools' record management. Academic institutions generate and invest in huge databases and data repositories that are able to store large amount of students-related data resulting in the problem of information (data) overload but knowledge starvation. In addition, the exponential growth of educational data and inefficiency of traditional exploratory techniques on these academic databases are major issues plaguing educational institutions (Bala and Ojha, 2012). Some of the sources of educational data are e-learning resources, internet connectivity, databases from student information systems, enterprise portals and instrumental educational software (Romero and Ventura, 2013).

Educational Data Mining (EDM), a sub-domain of data mining, is a trending discipline that focuses on the application of various methodologies, tools and algorithms, in the exploratory, graphical and intelligent analysis of educational data repositories for the discovery and extraction of new structures, which in turn will help

understand, predict and improve students' academic performances (Jacob, et al., 2015). EDM is aimed at providing an understanding of how students learn; and also identify the aspects that can improve learning and other educational activities (Villanueva *et al.,* 2018). EDM processes are used to provide real-time feedback exchange, or improving the learning management thereby enhancing the students' learning processes. EDM provides a conduit for lecturers/teachers/instructors to investigate, monitor and take students-centred actions aimed at improving students' learning processes. Thirdly, EDM is used to measure lecturers and students' experiences and approaches to learning. In this paper, EDM is applied to mine students' academic performance (AP) data, discover and extract hidden information and knowledge capable of promoting and supporting all facets of effective decision making relating to students' AP. EDM applications have gained much relevance recently because of its fundamental value in decision making and have become a pivot of educational institutions and other academic or professional bodies supporting any form of teaching. Its techniques have been introduced into new fields like statistics, databases and knowledge bases, Artificial Intelligence (machine learning, pattern recognition, computational intelligence) (Baradwaj, and Pal, 2011). Machine learning techniques such as decision trees, neural networks, naïve bayes, k-nearest neighbour, self-organizing maps (SOM), cluster analysis and many other approaches are employed to drive EDM processes. Using these techniques unknown but useful pieces of knowledge — association rules, classes and clusters, trend, relationships, models and so on, can be identified and extracted and thereafter used for descriptive and prescriptive purposes. For example, predictions regarding information on students' enrolment, evaluation of teaching methods and tools, identification of at-risk students through forecast of students' performance, discovery of students' performance(s) that are significantly different from the rest of data and partitioning students based on academic performance and so on (Baradwaj and Pal, 2011).

Cluster Analysis (CA) is an important task in data analysis. It is aimed at revealing implicit structures that were hidden, interesting patterns and relations from datasets, and then adapting the extracted information to the analysis task to ease comparison, interpretation and relationship assessment (Sacha, *et al.,* 2018). CA is tightly entwined with computations techniques, interactions and visualization to meet the requirements and expectations of modern real-world analysis problems (Sacha, et al., 2016, Sacha et al., 2017). Visualization is a coherent and compact graphical representation that reveals and communicates the complex details in patterns clearly, precisely, and efficiently (Sacha, *et al.,* 2018, Card, et al., 1998). Fundamentally, visualization provides an efficient means of gaining intrinsic and extrinsic details in data, as easy as possible, by analyzing, exploring, discovering, representing and communicating information and pieces of knowledge in well comprehensible form. There are many visualization tools used in different situations to convey different level of details. SOM is a special class of competitive neural network (NN) that is used extensively and successfully for pattern recognition, exploratory analysis, clustering, optimization and visualization of large databases (Eklund, et al., 2003). SOM is a suitable tool for any data type that can be represented by feature vectors including large, complex and multimedia data. It has a special property of effectively creating spatially organized internal representations of the various input features and their abstractions and capable of grouping objects according to similarity of relevant data features without changing the topology of the input feature space.

Several methodologies have been proposed for managing students' AP, yet the need for improved monitoring and management still lingers. In Inyang and Joshua (2013), students were clustered into weak, average and good clusters via k-means algorithm on dataset consisting of first year courses. Darcan and Badur (2012) investigate students' segments and profiles based on their various dimensions of academic abilities using cluster analysis. The work only considered dimension reduction of factors and clustered students with k-means but did not give visuals of the patterns of students' performance as well as relationship between the factors. In addition, the work reported in Inyang and Joshua (2013) and Darcan and Badur (2012) did not provide visualization of the attributes thereby making its interpretation arduous. This paper aims at discovering the number of clusters present in students' dataset, cluster students' AP data and discovers performance patterns. The rest of the paper is organized as follows, in section 2, a review of students' AP through EDM is presented while methodological workflow is described in section 3.0. Section 4 presents the Cluster visualization tasks while conclusions are drawn in section 5.

2. Students' Academic Performance and EDM

Academic institutions aim at imparting knowledge and skills, through teaching, training and mentoring, to students who pass through them, and using examination results to determine their AP levels. Students' AP involves an advancement of the students' knowledge and skills as measured by the Grade Point Average (GPA) and the gradual development of their personality and academic progress (Basri, et al., 2018). It is the desire of all students to earn high AP, since it is a significant indicator of positive outcomes which individuals, organizations and the society value. Students who are academically successful have high employability and productivity likelihood than those with poor grades. Poor AP of students is a major contributor to the high attrition rate and may also contribute to

the un-employability of students. The increasing demand for excellence in all areas of life has led to the need to evaluate, monitor and predict the possible AP outcome of students at an early stage in their studentship. Students' AP prediction is a desirable task in EDM and learning analytics, and plays a significant role in higher institutions of learning. The pressure of the educational managers and stakeholders— parents, guardians, teachers and school administrators, and the existing healthy competition among students and institutions have enabled the emergence of new strategies aimed at improving students' AP. These strategies include, extra classes for students, multiple admission modes and programmes, new teaching and learning methods and instructional strategies, motivational strategies for outstanding students and so on (Nyagosia, 2011).

GPA has been a globally adopted measure for assessing and monitoring students' learning outcome (Oyelade, Oladipupo, & Obagbuwa, 2010; Yadav & Singh, 2012;). The AP of students in their first year at higher institutions indicates the direction of the overall AP and contributes significantly on the Cumulative Grade Point Average (CGPA), which class of degree (CoD) depends upon (Shovon & Haque, 2012). Recommended learning management methodology involves the assessment of students from the inception of their studentship, gaining early feedback exchanges, monitoring the delivery and impact of support services, and providing information on the overall AP. The prediction of successful and unsuccessful students at an early stage provides an early categorization of students. This enables academic managers to concentrate on the bright students as well as develop remedial programs for the weaker ones in order to upgrade their AP while minimizing students' attrition. The procedure for the prediction of students' AP by educational planners is inefficient since they are based on statistical and database query approaches. Although statistical approaches are suitable for quantitative data, they fail to handle complex and noisy datasets efficiently because of their inability to perform pattern extraction (Inyang, 2012). Hence, the need for intelligent methodologies that can handle both large quantitative and ambiguous dataset features.

3. Methodological Workflow

The methodological workflow of this work as depicted in Figure 1, proceeds in the following steps; dataset collection and pre-processing, k-means and cluster validity analysis and SOM visualization. The pre-processing involves attribute selection, data categorization and exploratory analysis.

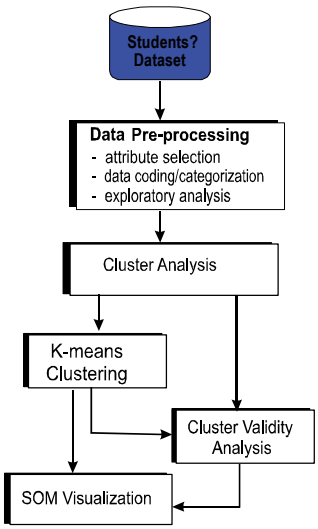

Figure 1. Workflow of the data mining methodology

K-means algorithm provided a means of partitioning the dataset into already known number of clusters (k=5) while cluster validity is performed using 19 clusters and four distance measures. SOM output is visualized in an hexagonal topological grid with feature component planes visualization. The detail description of each step is given in the following sections.

3.1 Dataset Description

The dataset used for clustering and classification of students' APs consists of nine sets (2004- 2013) of Bachelor of Science degree, Computer Science gradaunds of a university in Nigeria. It has three categories of academic

and demographic attributes — pre–admission indicators —Unified Tertiary Matriculation Examination (UTME) scores, post-UTME scores, demographic indicators —age, sex, and residential location and entry mode (EM) and post-admission attribute —performances of students at the completion of first year, graduation status, CoD. The UTME and post-UTME are the entrance examinations for prospective students of tertiary institutions in Nigeria. While UTME is organized by Joint Admission and Matriculation Board (JAMB), each university conducts its post-UTME. Scores that are earned in every course, correspond to grades 'A', 'B', 'C', 'D' or "F". Currently, a major requirement for graduation is a minimum of 'D' grade in all the compulsory courses prescribed for students in any programme. Upon satisfying the requirements, students graduate with a CoD – first class, second class upper, second class lower, third class and pass, depending on their graduating CGPA. Any student, who exceeds the specified period of programme on account of poor AP, is said to *'spill'*; such students are known as 'spillover*"* students. This paper aims at discovering unknown and important relationships and patterns resulting from students' AP using some selected demographic attributes, performances in UTME and post-UTME, EM and CoDs. Linguistic terms *"excellent", 'very good'*, *'good", 'fair",* and *'fail'* provide description of grades derived from scores, while '*very young'*, *'young'* and *'mature'* categorize age. *'graduated'*, *"voluntary withdrawal"* and *'spillover'* refer to students' status after the expiration of the stipulated period of a program.

3.2 Attributes Selection and Pre-Processing

Attribute selection involves the identification of the target vector and selection of the subsets or the input indicators on which the knowledge discovery process rely on. The input feature space comprises eight hundred and forty-six (846) observations with various factors that may affect AP of students. Data pre-processing task filtered redundant or irrelevant attributes from the original data and also categorizes textual attributes by converting them into numeric codes. The summary of the attributes and codes is presented in Table 1 while the number of students in each CoD at the end of the stipulated year in the university is presented in Table 2.

Table 1. Description of AP factors and their values.

S/N	Indicators	Description	Codes
1	Entry Mode	UTME	1
		Direct Entry	2
		On-campus	1
2	Residential Category	Off-campus	2
		Excellent	5
3		Very Good	4
	Scores_Courses	Good	3
		Fair	2
		Fail	1
4	P_UTME	High	3
		Average	2
		Low	1
		Very young	1
5	Age	Young	2
		Mature	3
		First class	5
		Second class upper	4
6	GPA	Second class lower	3
		Third class	2
		Pass	1
7	Sex	Male	1
		Female	2
		Ist Class	5
		2nd Class Upper	4
8	CoD	2nd Class Lower	3

	3rd Class	2
	Pass	1

The exploratory analysis gives a clear understanding of how the attributes are distributed in the dataset. For example, students' categorization based on sex yields 460 female and 386 male students. In CSC 111, 220 failed, 198 had fair grades while 60 had very good grade. From the dataset collected, 8.51% (72) of the students made second class upper. Two hundred and sixty six students, (266) belong to the pass CoD category which some in this category migth have dropped out. None of the students made a first class in the entire dataset collected.

Table 2. Distribution of students by CoDs in students' dataset

Cluster Number	**Description**	**Students' dataset**	
		Number of instances	**Proportion (%)**
1	Ist Class	0	0%
2	2nd Class Upper	72	8.5
3	2nd Class Lower	234	27.7
4	Third Class	274	32.4
5	Pass	266	31.4
Total		**846**	**100**

4. Cluster Visualization Analysis

The visualization of data mining results provides quick and clear details of the implicit complex interaction mechanisms and presents the discovered relations between the data that were hitherto incomprehensible although captured by the human visual system (Castillo-Rojas & Vega-Damke, 2017). In this paper, the visualization analysis is carried out in three stages; cluster discovery and validity analysis, k-means clustering and SOM clustering.

4.1 Cluster Discovery and Validity Analysis

Cluster Analysis (CA) is the process of discovering instances in a given dataset that can naturally group together and can be used to partition the dataset into classes. The objective of CA is to identify and describe classes within a dataset based on a well specified distance measure, so that the instances in the same class are similar to each other, while different to members of other classes (Maimon & Rokach, 2010). In some datasets, the actual number of clusters are known through the existence of natural divisions or the actual number of cluster are known a priori (for example, the students AP dataset used for this work); while others do not contain information on the natural divisions, or the natural divisions are unknown (Ekpenyong and Inyang, 2016). Cluster validation is an analytic process of assessing the quality of clustering solutions by finding the number of clusters that best satisfy the structure of the dataset without any priori class knowledge. There are three numerical similarity measures that are applied to assess various aspects of cluster validity; external index — used to evaluate the extent to which cluster labels match an external class label; internal index—examines the goodness-of-fit of the resultant structure without depending on external source of information; and relative index — produces a value from compare two different clustering solutions. Silhouette criterion, sum of squared errors, rand index, and entropy are some of the commonly used cluster validation metrics (Lyakh et al, 2012).

The silhouette criterion value (SCV) is an internal index that assesses the extent of cohesiveness between a point and other points within the same cluster, and the degree of separation between objects in different clusters. Equation 1 defines SCV for the *ith* data item, S(i), as follows (Rendón *et al.*, 2011):

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad - - - - - - - \quad 1$$

where $-1 \leq S(i) \leq 1$, a(i) is the mean distance between the *ith* object and other objects within a cluster, and b_i is the

mean distance between the *ith* object in a given cluster and objects in other clusters. The SVC closer to 1 depicts accurate cluster classification while values closer to -1 indicate poorly or incorrectly classified results. SCV hybridizes cohesion (degree of closeness of data-points within a cluster) and separation (degree of distinctiveness or well-separateness of points in one cluster is from another). It allows each data item, cluster and clustering task to be assessed by maximizing its value (Kodinariya, and Makwana, 2013). Liu et al. (2010) and Sivogolovko and Novikov (2012) identified that the silhouette performs well on a variety of data types irrespective of structural variations, noise and skewed distributions, and performs very well in partitioning and density-based approaches. This paper adopts silhouette criterion in the assessment of clusters by comparing pairs of objects between and within cluster distances (Liu et al., 2010, Liu and Sethuraman, 2013, Ekpenyong and Inyang, 2016). The optimum cluster number and cluster validation was based on experiments driven by k-means algorithm with SCV as the distance measure.

4.2 K-Means Clustering Analysis

K-means is a partitional algorithm that divides any set of data-points into disjoint clusters such that every object in the dataset belongs to only one cluster. This grouping is done on the basis of minimizing the sum-of-squared distances between objects and their respective centroids. The ease of use, simplicity and satisfactory performance across a wide variety of datasets was the basis for choosing k-means algorithm (DeFreitas and Bernard, 2015). Given the students' dataset comprising 846 objects and fives CoDs, k-means algorithm was performed by partitioning the dataset into a fixed number of clusters and thereafter, searching for the optimum number of clusters that best describes the structure of the dataset. In each phase, the centroids were randomly initialized while constructing k($n \geq k$) partitions. In the first phase, k was fixed at 5 (k=5) according to the number of CoDs and the results presented in Figure 2.

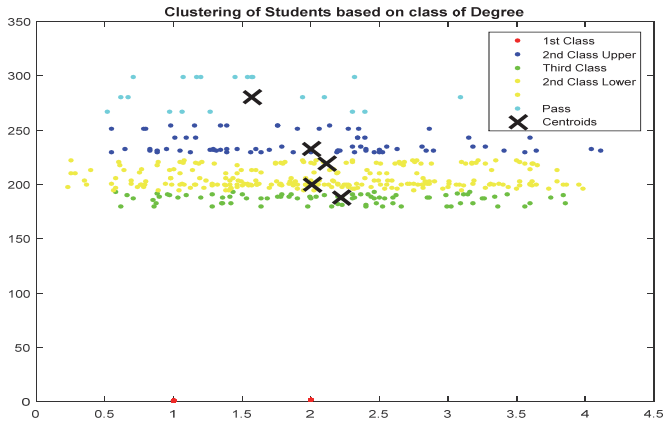

Figure 2. k-means clusters of students based on CoD

As depicted in Figure 2, each CoD has members and centroid of each cluster formed is marked. K-means discovers only two members for cluster 5 (first class) that was not originally present in the dataset. This means two members of cluster 4 (second class upper) were actually qualified to have been awarded 'first class' degree. The other clusters have significant number of members, with second class lower having the highest concentration of data-points as well as similar exchanges of members (especially members close to cluster boundaries) when compared with membership structure in the original students' dataset. Cluster labels could therefore be assigned to each corresponding record in the students' performance dataset, as target variable for supervised learning. Silhouette plot for the 5-clusters is shown in Figure 3. As shown in Figure 3, although the clustering solution is compact, they are not well-separated based on SCV and also cluster 5 is sparsely populated and this calls for cluster validity, which was performed with $2 \leq k \leq 19$ iterations. The performance of each cluster was assessed with four distance measures (squared euclidean, cosine, correlation and cityblock). The SCV in the various numbers of clusters and corresponding distance measures is presented in Table 3 and Figure 4.

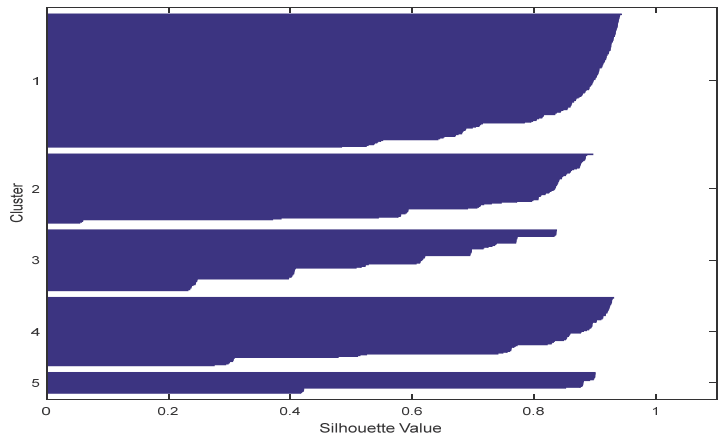

Figure 3. Silhouette graph of students' CoD (k=5) clusters

Table 3. SVC on various cluster numbers and distance metrics

No. of Clusters	Correlation	City block	Sqeucid	Cosine	**Average**
k=2	0.6517	0.2192	0.4426	0.6491	0.49065
k=3	0.6718	0.3497	0.5747	0.6611	0.570825
k=4	0.6887	0.3546	0.5220	0.6773	0.571300
k=5	0.6411	0.2622	0.5529	0.689	0.519475
k=6	0.6412	0.3002	0.5948	0.66537	0.550393
k=7	0.6396	0.2745	0.5153	0.6402	0.5174
k=8	0.6281	0.2644	0.5249	0.6265	0.510975
k=9	0.6400	0.2978	0.5523	0.6389	0.53225
k=10	0.6403	0.3277	0.5148	0.6183	0.525275
k=11	0.5973	0.2287	0.4800	0.5822	0.47205
k=12	0.6253	0.1840	0.5569	0.5674	0.4834
k=13	0.5614	0.2600	0.5052	0.5634	0.4725
k=14	0.6043	0.2170	0.5440	0.5119	0.4693
k=15	0.4922	0.2484	0.5592	0.5518	0.4629
k=16	0.5144	0.3233	0.5478	0.5076	0.473275
k=17	0.5071	0.2764	0.5522	0.5101	0.46145
k=18	0.5063	0.2684	0.5803	0.5315	0.471625
k=19	0.5162	0.3267	0.5276	0.4825	0.46325
k=20	0.4928	0.2694	0.6122	0.4788	0.4633
Average	**0.591726**	**0.275911**	**0.538232**	**0.585735**	

The SCV on varying cluster numbers generally decreases as the number of cluster increases, except in few instances. The top four performing numbers of clusters are 3, 4, 5 and 6 clusters with average SCVs of 0.570825, 0.571300, 0.519475 and 0.550393 respectively. The rank of clusters numbers also reveals the least performing number of clusters (19 and 20) with 0.46325 and 0.4633 as average SCV respectively. However, in terms of distance metric, correlation performed best with 0.591726 as the average SCV closely followed by cosine with a weight of 0.585735. In all four distance measures, the optimal number of clusters falls at 4, followed by six (6) clusters. The density of cluster 1— where many members are at the boundary or at the verge of dropping out, may account for the extra cluster when compared to the original dataset. However, since 4 clusters yielded the highest SCV, it means that this is the actual number of clusters existing in the dataset. In terms of similarity measure, correlation measure performed best while cityblock exhibited the least performance distance metric.

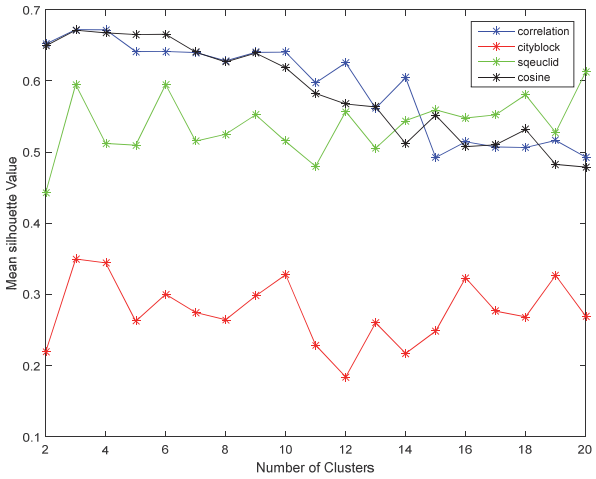

Figure 4. Comparative plots of silhouette values on number of clusters and distance measures

4.3 SOM Visualization

SOM algorithm is used to explore partitions and visualize students' dataset based on its distinctive support for data clustering, vector quantization, dimension reduction, and cluster visualization capabilities (Bernard, et al., 2011, Kohonen, et al., 2001). SOM was applied to identify structures and classify the students' dataset into the segments with similar performance characteristics. It was implemented in Matlab 2015a programming environment to integrate computation and rich visualization in four basic stages; initialization, competition, cooperation and adaptation. Dimensionality reduction was achieved during pre-processing, and involved scaling of data-points to minimize the influence of data points with high variance on other variables. Weight vectors were initialized by randomly assigning small values (set around 0) to nodes, as all instances of the dataset were processed, although an instance of a data-point may be processed more than once. An unsupervised batch bias algorithm called *trainbu,* which updates weights and biases after passing all the features into the SOM network. It was implemented through the rough and fine training phases; the rough training phase spanned a maximum of 1000 epochs and decremented the neighborhood radius and learning rate gradually from 5 to 2 and 0.5 and 0.1 respectively. This is to ensure global order at the commencement of training while local modifications to the map's model vectors became progressively specific as the radius reduces to zero. In the fine training phase, the learning rate was maintained at 0.2 with a maximum iteration of 500 epochs. Figure 5 describes the topology of the SOM model adopted for this paper.

The concentration of neurons in the output layer was set to 5x5 (as obtained from the CoDs), the links weight vector $(h_1^1, h_2^1, ..., h_n^j)$ of the links has the same dimension as the input feature vector and consists of prototypes linked with each node in the network. The input vector consists of feature representatives obtained from principal component analysis (PCA). Thereafter, the best performing distance measure for SOM analysis —correlation distance criterion was utilized to select the best representative (centroids) of the students' dataset feature within each cluster. Figure 6 shows Universal matrix (U-matrix) for the SOM component plots in an 8x13 hexagonal grid topology.

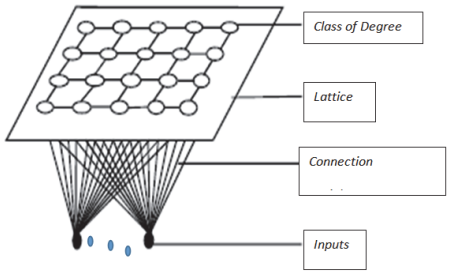

Figure 5. Structure of SOM Model

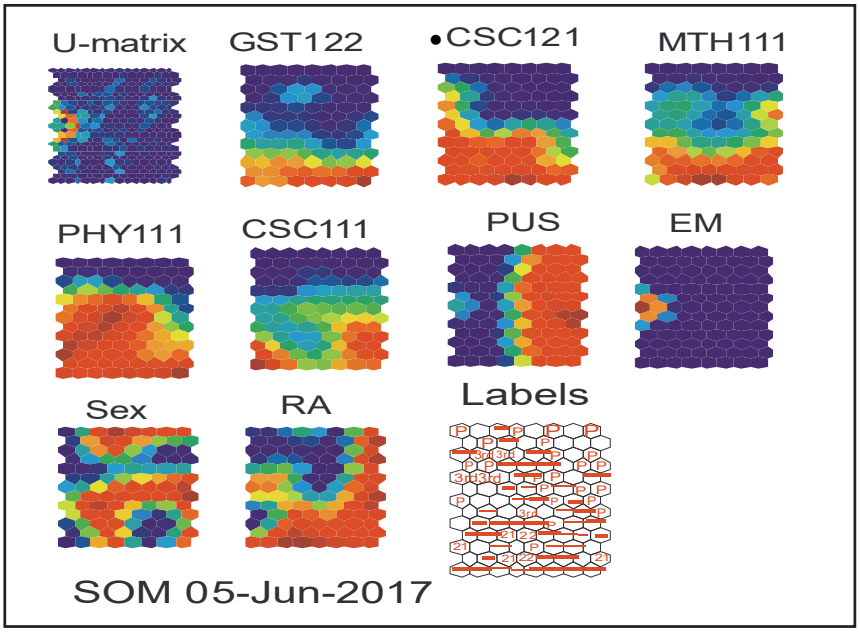

Figure 6. U-matrix and component planes visualization of significant AP features

The u-matrix is a density based graphical representation of the distances between neurons in the input dimension space. The darker blue colours indicate the lowest values in the data while the values increase as the colours get to the orange. High values of the U-matrix indicate a cluster border; uniform areas of low values indicate clusters themselves and reveal four disjoint clusters. The labels provide a clear understanding of the u-matrix us textual codes. The input feature component planes are representatives of the input features; GST 122, CSC 122, PHY111, CSC 111, RA (residence category) and PUS (Post-UTME score) and Sex. The GST 122 plane depicts a high membership in cluster 2 (third class) and cluster 5 (pass students). This is shown by the dark blue of the neurons around. This is a strong indication that GST 122 is a strong determinant of academic success in the case study programme. The CSC 121 feature plane reveals more neurons in the pass CoD and an equal number of second-class members in the map. This justifies the relevance of CSC 121 in the final performance of a student in computer science programme. PHY 111 shows few members belonging to pass CoD and a higher number of 'second class upper' students in its component plane. PUS had almost a balance in the third class and 'second class upper' students. The feature map planes reveal five groups of features based on their similarity. MTH 111, CSC111 and PHY 111 are highly correlated and belong to the same group while a similar pattern of neurons is observed in GST 122 and CSC 121 planes. The other features depict unique patterns, however RA and Sex had significant representatives in all the CoDs of the dataset while PUS had majority of points in only two clusters.

Figure 7 is a 9×9 sub-plot —scatter plots depicting the relationship and correlation between each input component and every other component in the dataset. The diagonal represents the histogram of the respective features while the upper triangles are the plots of the actual data-points. The lower triangle consists of map prototypes plots. The histogram shows that the scores in the course are similar and densely concentrated while the demographic attributes are sparsely distributed.

Figure 7. Correlation and relationships between students' AP indicators

The prototype plots of each indicator with others show a similar and related trend except with students' entry mode (EM). This implies that students with different EMs perform differently. In other words, impact of EM on students' performance in a particular course varies with other courses. While students' performance in other attributes are related and highly correlated — that is, any student who performs well in any of the courses will likely perform well in other courses. A significant drift in the performance in any of the courses will also be noticed in the other courses. The calibrated SOM visualization presented in Figure 8, gives an 8×13 hexagonal grid view of the number of clusters discovered by SOM and showing the four distinct clusters formed from the dataset. The distribution of neurons in each cluster gives the proportion (size) of each cluster. A mapping of the calibrated SOM map's cluster size to the original dataset shows that brown and yellow neurons (2nd Class Lower and third class) represents 33.7% each, deep blue (2nd Class Upper) accounting for 25% and cyan neurons (first class) with 7.7% of the map grid.

Figure 8. SOM calibration of the students' clusters based on CoD

Table 4. SOM map topology of discovered clusters

Cluster Name	Number of Neurons	Proportion (%)	Number of Instances
Ist Class	8	7.7	65
2nd Class Upper	26	25	212
2nd Class Lower	35	33.7	285
Third Class	35	33.7	285
Total	104	100	846

The map topology is given in Table 4. This confirms the result of cluster validity analysis described in section 4.1. As shown in Table 4, the proportion of the students in each cluster of SOM is greater than the corresponding cluster in the students' dataset. This reveals that some students that would have remained in their expected clusters or move to better clusters failed to do because they were not discovered at an early stage of the academic studies. The deviation in each of the clusters shows that 7.7% (65 students) of the students who were potential first class products did earn second-class upper degree. About 16.5% amounting to 140 students who were not monitored moved from cluster 2 into cluster 3. However, SOM cluster 1 (third class) had 30.1% of its member that were at-risk of either graduating without a degree (pass degree or without degree), spend extra year(s) in school or dropout. A summary of the deviation of the discovered clusters from the original clusters of the dataset is given in Table 5 while graphical distribution of the students in both dataset across the CoDs is given in Figure 9.

Table 5. Comparative Topology in SOM Clusters and Students' Dataset cluster

Cluster Number	**SOM Clusters**	**Students' dataset Clusters**	**At-Risk students**	
	Proportion (%)	**Proportion (%)**	**Proportion (%)**	**Number of students**
Ist Class	7.7	0	7.7	65
2nd Class Upper	25	8.5	16.5	140
2nd Class Lower	33.7	27.7	6	51
Third Class	33.7	63.8	30.1	255
Total	**100**	**100**	**60.3**	**510**

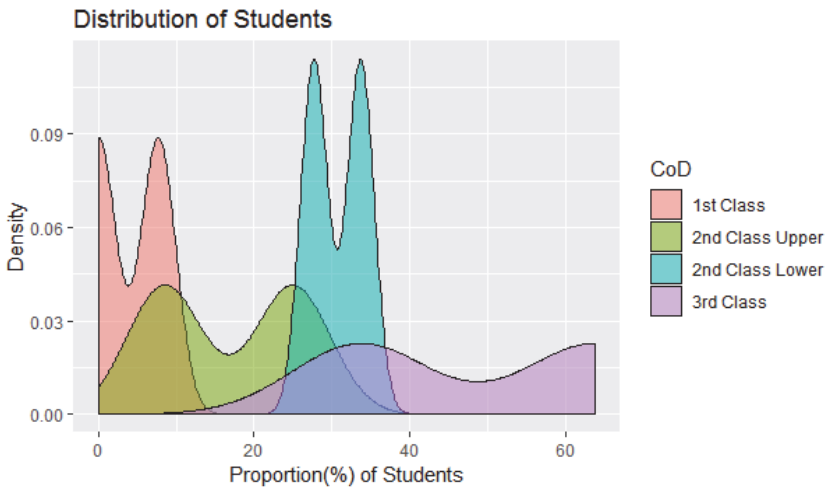

Figure 9. Density plot of distribution of students in both dataset across CoDs

As shown in Figure 9, the bandwidth of the SOM dataset and original students' dataset is 6.63 and 15.46, respectively. This implies that the distribution of the SOM is closer to the normal distribution than a larger bandwidth of the original students' dataset. There are two peak values in each cluster plot; the left hand peak represents the students' dataset while the right hand peak depicts SOM dataset values. The graph shows that all the four clusters overlap with at least one other cluster, which implies that some members of a cluster may also be qualified to belong to another cluster. For example, a greater portion of 2nd class upper cluster plot overlaps with SOM dataset portion of the density plot. Also, all the members of the 3rd class cluster from the students' dataset are also member of 2nd class lower and 2nd class upper clusters. This therefore suggests a proportion of 3rd class cluster were potential members of 2nd class lower and 2nd class upper clusters if they were monitored and managed properly. share It is therefore necessary to monitor students' AP in order to minimize their likelihood of been members of more than one cluster. This confirms the effectiveness of discovering at-risks students at an early stage to minimized wastages resulting from attrition, spending more than the stipulated duration of programme or poor CoD. The number of cluster discovered by SOM confirms the validity of the clustering solution obtained from k-means algorithm and provides a model for classification, predication students' AP and cluster validity analysis.

5. Conclusion

The large nature of student dataset has made it difficult to find patterns associated with students' AP using conventional methods and in turn making the management of students' AP an arduous task. AP management is of great importance, due to the impact of failure on individuals. Determining students who are likely to spend extra year in an institution or graduate with poor result at an early stage of their studies is of great importance. K-means was used for cluster validity analysis while SOM provided a means of clustering and visualizing the various attributes and clusters of AP. The system was implemented with Matlab 2015a programming environment and tested with nine sets of computer science students' dataset. The best performing number of cluster was 4 with *correlation* metric yielding the highest SCV of 0.5912. SOM provided a hexagonal grid visual of the dataset using component plane and scatter plots of each significant input attribute. The result shows that the significant attributes were highly correlated with each other except EM. This means that the impact of EM on the CoD varies with students irrespective of mode of admission. In addition, four distinct clusters were discovered in the dataset with SOM —7.7% belonging to cluster 1 (first class) and 25% belonging to cluster 2 (2nd class Upper). Cluster 3 and 4 had 35% proportion each. This number of clusters validates the results from k-means and further confirms the importance of early detection of students' performance, since the 7.7% of the students who were potential 1st class candidates eventually graduated with other CoDs because of lack of knowledge and programme to sustain them in that CoD. Moreso, the concentration of students with third class in the students' dataset was more than those discovered by SOM, meaning that members of cluster 3 drifted into cluster 4. This further validates the abolition of pass degree by the National University Commission and confirms SOM as a cluster validity tool. As further work, the labels from SOM will be associated with records in the dataset for association rule mining, supervised learning and prediction of students' AP.

References

Bala, M., & Ojha, D. B. (2012). Study of applications of data mining techniques in education. *International Journal of Research in Science Technology, 2012*(1), 1–10.

Baradwaj, B. K., & Pal, S. (2011). Mining Educational Data to Analyze Students' Performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 62-69. https://doi.org/10.14569/IJACSA.2011.020609

Basri, W, Alandejani, J., & Almadani, F. (2018). ICT Adoption Impact on Students' Academic Performance: Evidence from Saudi Universities. *Education Research International, 2018*(1), 1-9. https://doi.org/10.1155/2018/1240197

Bernard, J., Landesberger, T., Bremm, S., & Schreck, T. (2011). Multiscale visual quality assessment for cluster analysis with Self-Organizing Maps. In Proc. *SPIE Conference on Visualization and Data Analysis,* 1 –12, SPIE Press, 2011. https://doi.org/10.1117/12.872545

Card, S., MacKinlay, J., & Shneiderman, B. (1998). "Readings in Information Visualization: Using Vision to Think". Morgan Kaufmann.

Castillo-Rojas, W., & Vega-Damke, J. (2017). Visualization Proposal for Cluster Models with Multidimensional Data. Proceedings of the 21st World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2017), 120-125.

DeFreitas, K., & Bernard, M. (2015). Comparative performance analysis of clustering techniques in educational

data mining. *International Journal on Computer Science and Information Systems, 10*(2), 65-78.

Durkaya, M., & HÜSNÜOĞLU, M. (2018). The Role of Education in Employment. *Journal of Social Sciences and Humanities Researche, 19*(41), 51-70. https://doi.org/10.1080/03075070802457082

Eklund, T., Back, B., Vanharanta, H., & Visa, A. (2003). Using the self-organizing map as a visualization tool in financial benchmarking. *Journal of Information Visualization, 2003*(2), 171-81. https://doi.org/10.1057/palgrave.ivs.9500048

Ekpenyong, M. E., & Inyang, U. G. (2016). Unsupervised Mining of Under-resourced Speech Corpra for Tone Feature Classification. International Joint Conference on Neural Networks (IJCNN), 2374 – 2381. https://doi.org/10.1109/IJCNN.2016.7727494

Inyang, U. G. (2012). Development of Knowledge Discovery System for Oil Spillage Risks Pattern Classification. *Journal of Artificial Intelligence, 3*(4), 73-86. https://doi.org/10.5430/air.v3n4p77

Inyang, U. G., & Enobong, E. J. (2013). Fuzzy Clustering of Students' Data Repository for At-Risks Students Identification and Monitoring. *Computer and Information Science.* https://doi.org/10.5539/cis.v6n4p37

Jacob, J., Kavja, J., Kotak, P., & Puthran, S. (2015). Educational Data Mining Techniques and their Applications. International Conference on Green Computing and Internet of Things (ICGCloT). https://doi.org/10.1109/ICGCIoT.2015.7380675

Kodinariya, T., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies, 1*(6), 90-95

Kohonen, T., Schroeder, M. R., & Huang, T. S. (2001) Self-Organizing Maps. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd ed.

Liu, S., Matzavionos, A., & Sethuraman, S. (2013). Random Walk Distances in Data Mining and Applications. *Advance Data Analysis and Classification,* 83-108. https://doi.org/10.1007/s11634-013-0125-7

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu J. (2010). Understanding of Internal Clustering Validation Measures. In Data Mining (ICDM), 2010 IEEE 10th International Conference on. Sydney, New South Wales, 1–6.

Lyakh, Y., Gurianov, V., Gorshkov, O., & Vihovanets, Y. (2012). Estimating the Number of Data Clusters Via The Contrast Statistic. *Journal of Biomedical Science and Engineering,* 95-99. https://doi.org/10.4236/jbise.2012.52012

Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook, 2nd ed. Springer Science+Business Media, Inc. Edited by Oded Maimon and Lior Rokach Tel-Aviv University, Israel. ISBN 978-0-387- 09822-7, 2010.

Nyagosia, P. O. (2011). Determinants of Differential Kenya Certificate of Secondary Education Performance and School Effectiveness in Kiambu and Nyeri Counties, Kenya. Kenyatta University. Retrieved April 25, 2018 from http://irlibrary.ku.ac.ke/bitstream/handle/123456789/3009/Nyagosia,%20Patrick%20Ogecha.pdf?sequence=3

Oyelade, O., Oladipupo, O., & Obagbuwa, I. (2010). Application of K-means Clustering Algorithm for Prediction of Student's Academic Performance. *International Journal of Computer Science and Information Security, 7*(1), 292-295.

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus External Cluster validation indexes. *International Journal of Computer and Communication*, 28-34.

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining Knowledge Discovery, 2013*(3), 12–27. https://doi.org/10.1002/widm.1075

Sacha, D., Kraus, M., Bernard, J., Behrisch, M., Schreck, T., Yuki A., & Keim, D. (2018). SOMFlow: Guided Exploratory Cluster Analysis with Self-Organizing Maps and Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 120-130

Sacha, D., Sedlmair, M., Zhang, L., Lee, J. A., Weiskopf, D., North, S. C., & Keim, D. A. (2016). Human-Centered Machine Learning Through Interactive Visualization: Review and Open Challenges. Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.

Sacha, D., Zhang, L., Sedlmair, M., Lee, J. A., Peltonen, J., Weiskopf, D., North, S. C., & Keim, D. A. (2017). Visual interaction with dimensionality reduction: A structured literature analysis. IEEE Trans. *Visualization and Computer Graphics, 23*(1), 241–250. https://doi.org/10.1109/TVCG.2016.2598495

Shovon, M., & 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, 3*(8), 146-149.

Sivogolovko, E., & Novikov, B. (2012). Validating cluster structures in data mining tasks. In Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT '12. EDBT-ICDT '12. New York, USA: ACM, p. 245-205.

Vanthienen, Jan, & Witte, Kristof De (2017). Data Analytics Applications in Education. Taylor and Francis. Retrieved Dec 27, 2018, from https://www.researchgate.net/publication/320226279_Data_Analytics_Applications_in_Education

Villanueva, A., Moreno, L. G., & Salinas, M. J. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review, 33*(2018), 235-262. Retrieved from http://greav.ub.edu/der/

Yadav, R. S., & Singh, V. P. (2012). Modeling Academic Performance Evaluation Using Fuzzy C-Means Clustering Techniques. *International Journal of Computer Applications, 60*(8), 15-23.