

Fuzzy Clustering of Students' Data Repository for At-Risks Students Identification and Monitoring

Udoinyang G. Inyang¹ & Enobong E. Joshua²

¹ Department of Computer Science, Faculty of Science, University of Uyo, Uyo, Nigeria

² Department of Mathematics & Statistics, Faculty of Science, University of Uyo, Uyo, Nigeria

Correspondence: Udoinyang G. Inyang, Department of Computer Science, Faculty of Science, University of Uyo, Akwa Ibom State, Nigeria. Tel: 234-80-3668-8711. E-mail: udoiinyang@yahoo.com

Received: July 02, 2013 Accepted: August 8, 2013 Online Published: August 21, 2013

doi:10.5539/cis.v6n4p37

URL: <http://dx.doi.org/10.5539/cis.v6n4p37>

Abstract

In educational data mining, identifying academic courses that contribute significantly to students' class of degree and predicting students' performances can help in the choice and improvement of intervention and support services for students whose performances are poor. Experience shows that graduates with weak class of degree find it difficult to gain employment, hence, the need to identify and group these at-risk students at an early stage of their academic career and then develop a plan to improve their performance. This paper identifies possible academic courses with significant contribution to academic performance and predicts students' graduating class of degree. 11Ants Model Builder provided a means for course rank analysis while MATLAB was the system development tool. Fuzzy c-Means (FCM) algorithm was used to partition students into *weak*, *average* and *good* clusters. Four (4) natural clusters of at-risk students were automatically identified with k-means algorithm. Results show that Sugeno-type inference system is best suitable for the provision of initial parameters for Adaptive Neuro Fuzzy Inference System (ANFIS) training of students' dataset. The results also prove the effectiveness of the combination of FCM, k-means and ANFIS in the classification of students based on academic performance and at-risk levels. The results will help educational managers monitor groups of students at the same level of performance, and those at the boundary of two classes of degree for the provision of informed counseling and intervention plans, to improve academic performance.

Keywords: students' performance, significant courses, fuzzy clustering, grade point average, at-risk students

1. Introduction

The increasing demand for information from educational planners has led to the existence of huge educational data repositories for use in the extraction of vital information. Chandra and Nandhini (2010) describes students' result repository as a large data bank of students' raw scores and grades in different courses enrolled for during their years of attendance in an institution. The data, which can be personal or academic, can be used to understand students' behaviour, assist instructors, improve teaching, evaluate and improve e-learning systems and many other benefits (Romero & Ventura, 2007). As huge amount of data is being collected and stored in databases, traditional statistical and database management techniques are no longer adequate for analyzing them (Kumar & Chadha, 2011). Although, statistical approaches quantify the inherent uncertainty that results when one tries to infer general patterns from a particular sample or an overall population, they lack the ability to handle large, complex, noisy and vague dataset. Furthermore, these approaches perform limited search during pattern extraction from databases, thereby producing incomplete and unreliable results, which cannot support effective decision-making (Fayyad, Piatetsky-Sapiro, & Smyth, 1996; Miller & Han, 2001). This gives rise to improved techniques and tools for automatic and intelligent analysis of huge data sets. One of such techniques is Data Mining (DM), which has attracted attention from researchers in many different fields such as database design, statistics, pattern recognition, machine learning, data visualization, biology, web applications, electronic commerce and so on (Piatetsky-Shapiro, 2007).

Cluster analysis is an unsupervised learning technique of DM, which, through the examination of relationships existing between data, describes and groups them into classes or clusters (Yang, 1993; Rao & Vidyavathi, 2010; Shovon & Haque, 2012). It is a process of grouping data objects into disjointed clusters so that the data in the same cluster are similar while data belonging to different clusters are different (Kumari, Sharma, & Gaur, 2012).

Many clustering algorithms have been introduced, all based on the principle of maximizing the similarity between objects in the same cluster and maximizing the dissimilarity between cases of different clusters (Han, Kamber, & Tung, 2001). Baboo and Priya (2013) identify two broad categories of clustering algorithms; Hard (k-means) clustering partitions the data into a specified number of mutually exclusive subsets. Hard clustering, forces data points, in any sample space, that have attributes of more than one cluster to a particular cluster. Fuzzy clustering, however, allow objects to belong to more than one cluster simultaneously, with different degrees of membership. Objects on the boundaries between several classes are not forced to fully belong to only one of the classes, but rather are assigned membership degrees in the range $[0, 1]$ indicating their partial membership in each classes (Kumar, Verma, & Sharma, 2010; Baboo & Priya, 2013; Zuviria & Deepa, 2013). Fuzzy C-Means (FCM), Possibilistic C-Means (PCM), Fuzzy Possibilistic C-Means (FPCM) and Possibilistic Fuzzy C-Means (PFCM) are some of the fuzzy clustering techniques. Clustering Analysis is a method of data description which provides a means of data analysis in various fields like machine learning, data mining, pattern recognition, image analysis and bio-informatics (Ghosh & Dubey, 2013). The FCM clustering is best model for students' academic performance modeling and serves as a good benchmark to monitor the progression of students in educational domain (Yadav & Singh, 2012).

Educational Data Mining, concerns with the extraction of useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of students' learning process (Chan, Chow, & Cheung, 2008; Romero, Ventura, & Garcia, 2008; Ngor, 2007; Castro, Nebot, & Mugica, 2007). Students' academic performance monitoring plays a very vital role in higher institutions of learning. Grade Point Average (GPA) is a common factor used by the academic planners to evaluate and monitor the progression of students (Sansgiry, Bhosle, & Sail, 2006; Oyelade, Oladipupo, & Obagbuwa, 2010; Yadav & Singh, 2012; Shovon & Haque, 2012). The academic performance of students during their first year at university is a turning point in the path of their educational performance and usually influences their Cumulative Grade Point Average (CGPA) significantly, which in turn affects their class of degree (Shovon & Haque, 2012). Based on this critical issue, grouping students into different clusters according to their performance has become a useful but complicated task (Oyelade et al., 2010). Yadav and Singh (2012) states that intelligence-level wise grouping is essential for maintaining the homogeneity of the group otherwise it would be difficult to provide good educational services to student population with highly diverse characteristics.

The main goal of fuzzy clustering analysis is to partition students into homogeneous groups according to their characteristics and abilities (Kifaya, 2009). Oyelade et al. (2010) demonstrated the combination of k-means algorithm and deterministic model provided in (Omolehin, Oyelade, Ojeniyi, & Rauf, 2005) in the prediction of students' academic performance. The system provided clusters of students and the overall performance of each cluster by cluster size. However, the system proposed in (Oyelade et al., 2010) lacked the facility to handle uncertainty characterizing academic performances. Yadav and Singh (2012) proposed a rule based Fuzzy Expert system for students' academic performance evaluation based on FCM Clustering algorithm, Fuzzy Logic (FL) and Regression analysis model. FL models lack self-learning and adaptive capabilities (Muzzammil, 2010) while Regression Models are effective where there are no multicollinearity among the predictor variables (Petraitis, Dunham, & Niewiarowski, 1996; Giovanis, 2010). On the other hand, Adaptive Neuro-Fuzzy Inference System (ANFIS) provides an intelligent way of reasoning and prediction, and performs better than regression models and other conventional statistical techniques (Aali, 2009; Muzzammil, 2010; Bisht & Jangid, 2011; Mahdavi & Khademi 2012; Giovanis, 2010). ANFIS is the widely used model in the studies of classification, estimation and prediction (Yayar, Hekim, Yilmaz, & Bakirci, 2011). This paper proposes a methodology based on the hybrid of FCM, k-means algorithms and ANFIS for the prediction of students' performances and classification of students based on performance and at-risk levels.

2. Overview of Fuzzy c-Means (FCM) Algorithm

The FCM algorithm is an iterative algorithm that generalizes the hard c-means algorithm to allow any point partially belong to multiple clusters (Kumar et al., 2010). The aim of FCM is to find clusters centers that minimize a dissimilarity function and then partition a finite collection of elements, $X = \{x_1, x_2, x_3, \dots, x_n\}$, into a collection of fuzzy clusters, $C = \{c_1, c_2, \dots, c_p\}$ with respect to some given criterion (Ekong, Onibere & Imianvan, 2011). The algorithm is implemented in the following steps (Inyang & Akinyokun, 2006; Kumar et al., 2010; Imianvan & Obi, 2011; Zuviria & Deepa, 2013).

- i. Set m, c , and ϵ , such that $m > 1$, $2 \leq c < n$, and $0 < \epsilon < 1$
- ii. Initialize $U = [u_{ij}]$ Matrix, $U(0)$
- iii. At k -step: calculate the centre vectors $C(k) = [c_i]$ with $U(k)$

$$C_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m} \quad (1)$$

iv. Update $U(k)$, $U(k+1)$

$$u_{ij} = \sum_{k=1}^p \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{-2}{m-1}} \quad (3)$$

with the following constraints:

$$0 \leq u_{ij} < 1, \quad \forall i, j$$

$$\sum_{j=1}^p u_{ij} = 1, \quad \forall i \quad (4)$$

v. if $\|U(k+1) - U(k)\| < \varepsilon$ then STOP: otherwise return to step (iii)

where,

- U - partition matrix
- u_{ij} - degree of membership of x_i in the cluster j
- x_i - the i th of d -dimensional measured data
- c_j - the d -dimension centre of the cluster
- ε - termination criterion
- k - maximum number of iteration steps
- m - the fuzzification parameter
- p - the number of clusters
- d - the dimension of the dataset

3. Methodology

3.1 Dataset Selection, Description and Preprocessing

The students' dataset used for the training and analysis, consists of six sets of Bachelor of Science, Computer Science graduands. The input variables of interest are performances of students in first year courses while the target variable is students graduating class of degree. Each student's performance in a course, measured by their score in the course, is an aggregation of Continuous Assessment (CA) score and the examination scores. CA score constitutes 30% of the total score while examination is 70%. The various grades associated with scores are described in Equation 5. Each grade is assigned a numerical value or point; grade 'F' attracts 0 point, 'E' is weighted 1, 'C' attracts 3 points while 'A' grade is weighted 5 points. The product of the point and Course Unit (Credit Hours) yields the Quality Point (QP) earned by a student in the course. The academic standing of any student is based on GPA, determined by dividing the total QP earned by a student in a semester by the total credit hours of courses registered for during the semester. The class of degree of any student depends on the graduating CGPA, which is in the range [0.00, 5.00].

$$g(x) = \begin{cases} "A" & \text{if } Score > 69 \\ "B" & \text{if } 60 \leq x < 70 \\ "C" & \text{if } 50 \leq x < 60 \\ "D" & \text{if } 45 \leq x < 50 \\ "E" & \text{if } 40 \leq x < 45 \\ "F" & \text{if } x < 40 \end{cases} \quad (5)$$

where $g(x)$ is the grade associated with score x .

Experience shows that students who perform below second-class lower division are regarded to have performed poorly; those with second-class lower division are considered as average students, while students with second class upper division and first class are rated as very intelligent. In this work, three clusters of students, *Weak*

(Third Class and Pass Degrees), *Average* (Second Class Lower Division) and *Good* (Second Class Upper and First Class Honours) were considered. A summary of the clusters and their respective CGPA is presented in Table 1.

Table 1. Clusters and clustering criteria of students

Cluster No.	Cluster Name	Class(es) of Degree	CGPA Range
1	Weak Students	Pass degree, Third Class	[0.00, 2.39]
2	Average Students	Second Class Lower Division	[2.40, 3.49]
3	Good Students	Second Class Upper Division, First Class Honours Division	[3.50, 5.00]

As shown in Table 1, all CGPAs < 2.40 are labeled *Weak* class of degree, CGPA in the range $[2.4, 3.49]$ is labeled *Average* class of degree while CGPA ≥ 3.49 is labeled *Good* class of degree. Students that failed to complete their studies on account of voluntary withdrawal and those who had missing result(s) in any of the courses were excluded from the dataset. The research relied on 496 complete records of Bachelor of Science (B.Sc.), Computer Science students admitted from 1999 to 2005 in University of Uyo, Nigeria. The sequence of steps adopted in the clustering of students' data repository is presented in Figure 1.

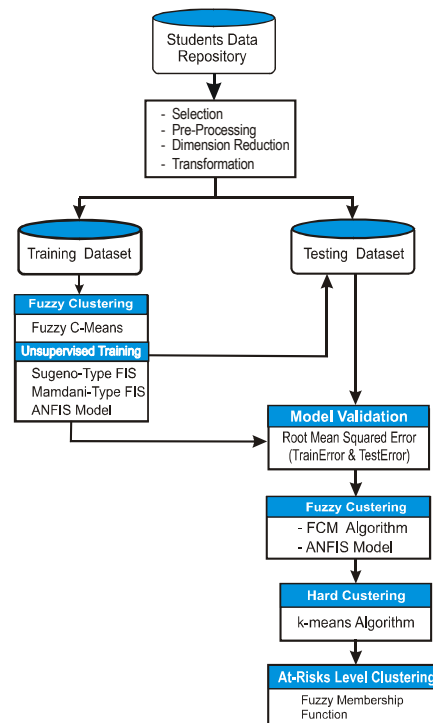


Figure 1. Methodology for fuzzy clustering of students data repository

3.2 Input Variables Rank Analysis

An attribute importance analysis is necessary to rank attributes based on their contribution to the target values and to reduce the size of input variables for prediction. It will also help to increase speed and accuracy of models in the prediction task (Paris, Affendey, & Mustapha, 2010; Inyang, Njungbwen, & Inyang, 2009; Saltelli, 2002). In the domain of educational data mining, identifying dominant courses can help improve the intervention and support services, at an early stage, for students who perform poorly in their studies. It will also help the instructors to concentrate in the teaching of courses with high rate of failure and which have significant contribution to academic performance (Chamillard, 2006). 11Ants offers a straightforward and effective means of estimating and testing a very large number of models and ensembles within the convenient and familiar framework of MS-Excel. 11Ants also has strengths in terms of wider range of algorithms, easy data preparation

tools and supports very large dataset (Inyang, 2011). 11Ants Model Builder was used to determine each course's influence on the class of degree. A rank of courses based on their importance is presented in Table 2.

Table 2. Course importance value and rank

S/N	Course Code	Importance	Rank
1	MTH 121	0.093	1
2	PHY 111	0.081	2
3	GST 122	0.076	3
4	CHM 111	0.075	4
5	MTH 122	0.073	5
6	CSC 111	0.065	6
7	CSC 121	0.061	7
8	GST 114	0.0593	8
9	BIO 111	0.0590	9
10	MTH 111	10	0.051
11	PHY 121	11	0.049
12	PHY 122	12	0.048
13	CHM 121	13	0.045
14	GST 112	14	0.044
15	PHY 112	15	0.042
16	GST 111	16	0.0394
17	GST 121	17	0.0388

The results depicted in Table 2 show that core departmental courses like MTH 121 (General Mathematics II), MTH 111 (General Mathematics I), CSC 111 (Introduction to Computer Science) and CSC 121 (Introduction to Computer Programming) have strong correlation with class of degree than university wide courses like GST 111 (Use of English II) and GST 112 (Nigerian People and Culture). GST 122 (Introduction to Philosophy and Logic), a university wide course, also proved significant. Courses that have importance value greater than 45% has a cumulative effect of 79% on the class of degree, hence are significant. Courses with weight less than 0.46 were noisy and insignificant; therefore pruned from the dataset. The dataset was randomly split into two; 446 records as the training dataset and 50 as the test dataset. Students' GPA derived from the identified significant first year courses were used for training and clustering of students. In this work, at-risk students are those students who perform poorly and are likely to graduate with a weak class of degree. Each student's at-risk level is determined by his/her degree of membership in the weak students cluster.

4. Fuzzy c-Means System and Results

This paper considers a method by which fuzzy membership functions are created for clusters of students based on GPA derived from the significant first year courses identified in Section 3.2 and the vector of degree of membership of students in weak students cluster. The system was developed in Matrix Laboratory (Matlab) with Ms Excel as the backend engine. The major components of the system are Knowledge Base (KB) and Inference Engine. The KB contains students' raw scores, grades and GPAs. It also contains fuzzy rules, linguistics values and fuzzy membership values. Two clustering algorithms; FCM and k-means algorithms, and Adaptive Neuro-Fuzzy Inference System (ANFIS) drive the Inference Engine. The FCM parameters were set as follows; $m=2$, $\epsilon=0.01$, and $k=200$. The arithmetic mean of all the data points was the initial cluster center. The performances of the FCM objective function while partitioning the training data set into three clusters is depicted in Figure 2.

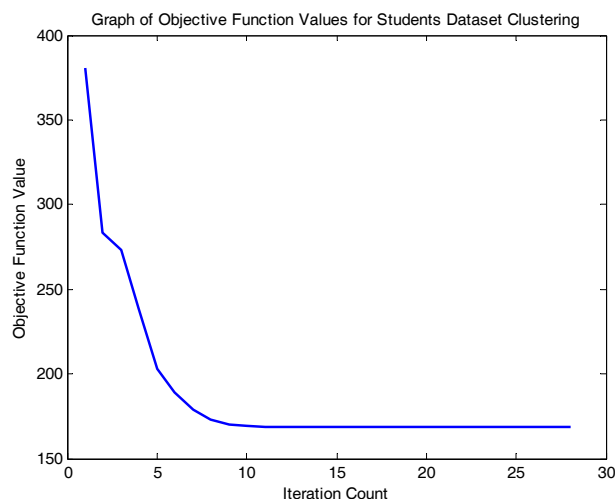


Figure 2. Graph of objective function values for students' dataset clustering

As shown in Figure 2, the initial value of the objective function is 362.22, which decreases to 168.44 at the 11th iteration. However, the change in the objective function value is very negligible between the 11th and 27th iteration without further change after the 27th iteration. The performance wise clustering of students in the training data is presented in Figure 3.

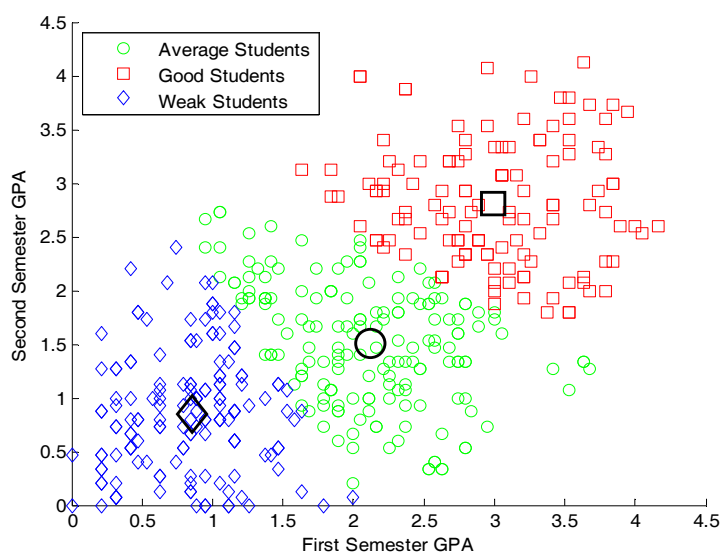


Figure 3. Graph of performance level clusters and cluster centre of students dataset

Figure 3 shows that, 126 students have highest membership degrees in weak students' cluster, 170 students are in Average students cluster while the Good students cluster has 150 students. The large symbols represent the final cluster centre obtained from the training. A summary of the cluster centres and numbers of students is depicted in Table 3. Fuzzy Inference Systems (FISs) structures were generated for ANFIS. Sugeno-type (fismat1) and Mamdani-type (fismat2) FISs were built by extracting rules that model the students dataset behaviour using membership functions for rules' antecedent and consequent parts. The third FIS (fismat3), a Sugeno-type was generated using subtractive clustering in determining the number of rules and antecedent membership functions; and linear least squares estimation method for determining each rule's consequent. The summary of the performances of these FISs on the dataset is presented in Table 4.

Table 3. Description of performance level clusters of students

Cluster Number	Cluster Name	Cluster Centre		No. of Students (Cluster size)
		First Semester GPA	Second Semester GPA	
1.	Weak Students	0.851	0.843	126
2.	Average Students	2.122	1.500	170
3.	Good Students	3.00	2.806	150

Table 4. Performance of FISs on the students dataset

S/N	FIS	Type	Training Error (trnRMSE)	Checking Error (chkRMSE)
1	Fismat1	Sugeno	0.4972	0.2842
2	Fismat2	Mamdani	0.5247	0.3897
3	Fismat3	Sugeno	0.4950	0.2923

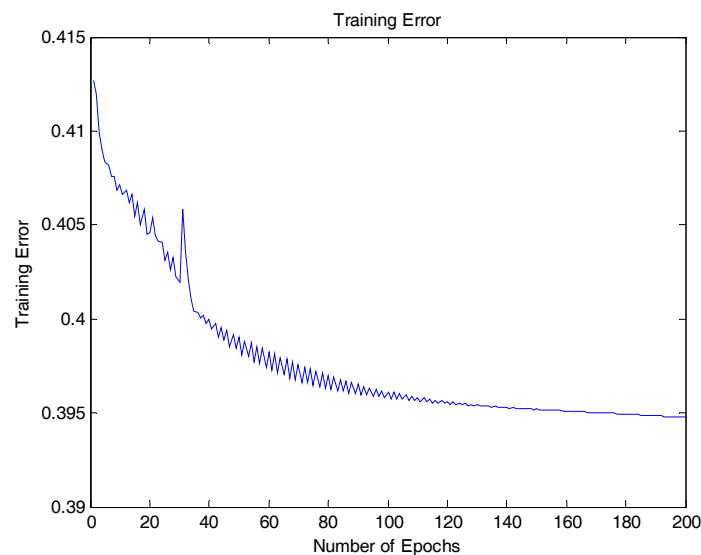


Figure 4. Graph of training error of FCM clustering system

The performance metric used was the Root Mean Squared Error (RMSE) for training (trnRMSE) and testing (chkRMSE) of the dataset as in (Hossain & Ahmad, 2012). The results show that Sugeno-type FIS has lower RMSE for both training and testing sessions, therefore better than Mamdani FIS. However, trnRMSE for Fismat1 is slightly higher than that of Fismat3, though with a lower chkRMSE of 0.2842. Fismat3 was chosen for the provision of initial parameters for training ANFIS. The properties of the ANFIS Model are presented in Table 5. The ANFIS training errors depicted in Figure 4 shows that the error settles at the 100th epoch as the ANFIS attempts to minimize the error.

Table 5. Description of training parameters for ANFIS

S/N	ANFIS Parameter	Values
1	Number of nodes	23
2	Number of linear parameters	9
3	Number of nonlinear parameters	12
4	Total number of parameters	21
5	Number of training data pairs	446
6	Number of checking data pairs	50
7	Number of fuzzy rules	3

5. Model Validation and Evaluation

The system was validated and evaluated using the test dataset. Figure 5 shows the plot of test data and predicted values while Figure 6 is the graphical representation of testing error. The circles in Figure 5 represent the predicted output while the lines represent the test dataset.

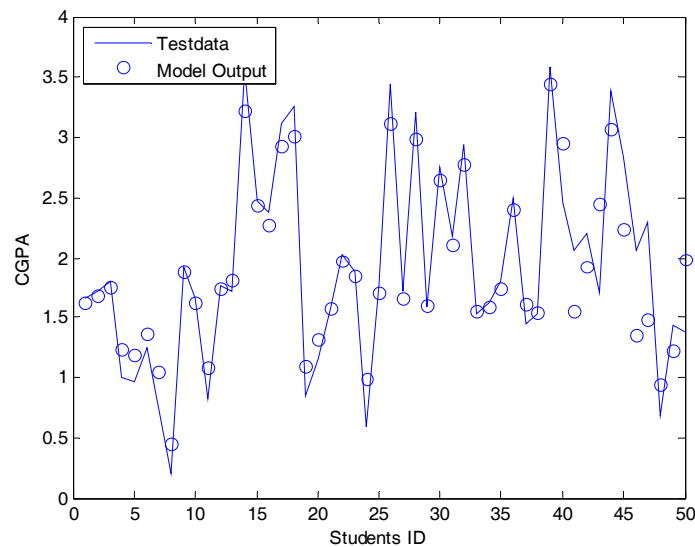


Figure 5. Graph of test dataset output and model output

The graph depicted in Figure 6 shows that the smallest value of the testing data error occurs at the 58th epoch, after which it increases slightly with an average step size value of 0.00042 even as ANFIS continues to minimize the error against the training data to the 200th epoch. The RMSE of the model on the test data is 0.2819, this shows that the system's performance is satisfactory and suitable for the prediction and clustering of students based on performance level. The results of FCM clustering of the Test dataset are presented in Table 6 and Figure 7.

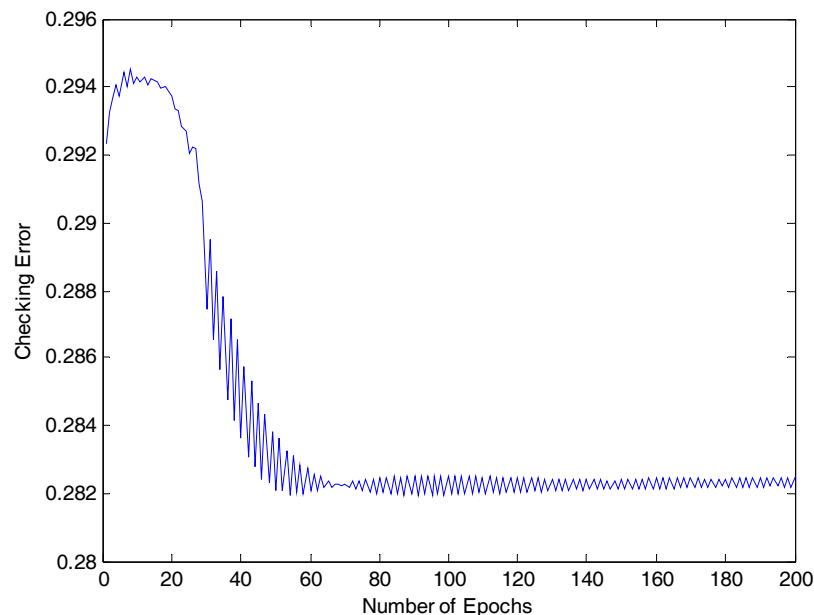


Figure 6. Graph of model testing error

Table 6. FCM degree of membership of students in each cluster

Student ID	Degree of Membership			Student ID	Degree of Membership		
	Weak Students	Average Students	Good Students		Weak Students	Average Students	Good Students
1	0.79	0.18	0.04	26	0.03	0.96	0.01
2	0.02	0.90	0.08	27	0.98	0.01	0.01
3	0.89	0.10	0.01	28	0.03	0.86	0.11
4	0.94	0.01	0.05	29	0.04	0.73	0.23
5	0.99	0.01	0.00	30	0.07	0.76	0.17
6	0.02	0.25	0.73	31	0.04	0.67	0.30
7	0.86	0.03	0.11	32	0.02	0.89	0.09
8	0.05	0.17	0.78	33	0.76	0.17	0.07
9	0.03	0.92	0.05	34	0.46	0.47	0.07
10	0.01	0.93	0.06	35	0.13	0.74	0.14
11	0.00	0.02	0.97	36	0.38	0.56	0.06
12	0.09	0.69	0.22	37	0.23	0.53	0.24
13	0.14	0.52	0.34	38	0.70	0.22	0.08
14	0.94	0.04	0.02	39	0.07	0.60	0.33
15	0.08	0.75	0.17	40	0.13	0.84	0.05
16	0.24	0.59	0.17	41	0.01	0.97	0.02
17	0.43	0.46	0.11	42	0.98	0.01	0.01
18	0.02	0.85	0.14	43	0.34	0.54	0.12
19	0.95	0.01	0.04	44	0.15	0.84	0.01
20	0.82	0.02	0.16	45	0.33	0.46	0.21
21	0.80	0.17	0.03	46	0.72	0.25	0.04
22	0.62	0.36	0.03	47	0.51	0.45	0.03
23	0.80	0.05	0.15	48	0.98	0.02	0.00
24	0.03	0.93	0.04	49	0.91	0.08	0.02
25	0.01	0.97	0.02	50	0.83	0.15	0.02

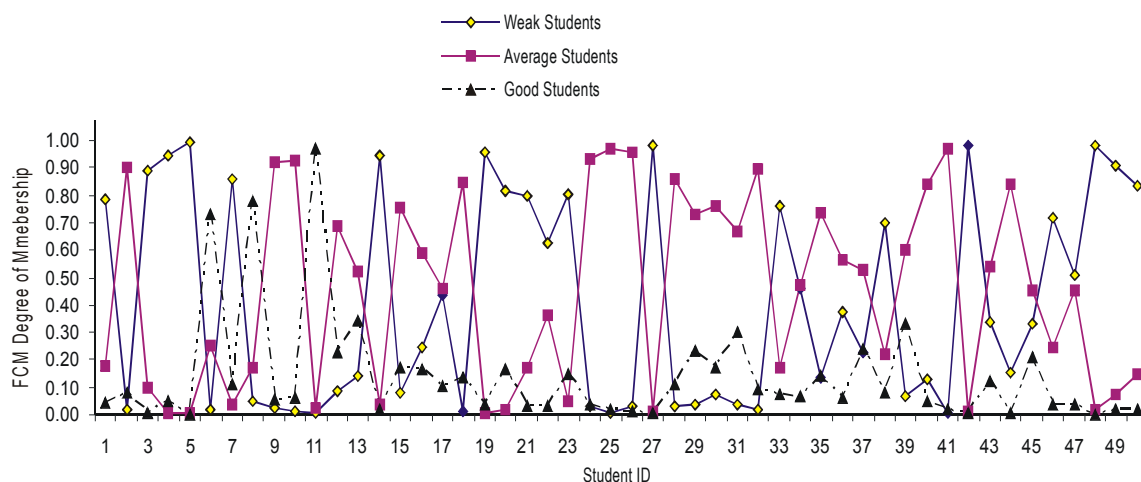


Figure 7. Graph of FCM Degree of Membership of Students in each cluster

The results presented in Table 6 and Figure 7 show that student number 1 belongs to the weak students cluster with a degree of membership 0.79 while student number 2 has 90% likelihood of graduating with second class lower degree (Average students cluster), 2% and 8% likelihood of weak and good students clusters respectively. In addition, students' number 4, 5, 7, 14, 19, 27, 42, 48 and 49 are at risk of attrition with weak students' membership value greater than 90%. These students may be advised to withdraw from the programme. Students number 17, 22, 34, 36, 43, 45 and 47 have competing degree of membership in two clusters. A plan for these students should be provided to either move them from a lower cluster to a better one or sustain them in the better cluster. For example student number 34 has a membership value of 0.46 in weak students cluster and 0.47 in average students cluster, therefore should be monitored for improvement required to sustain the student in the average cluster. Furthermore, student number 37 depicts 0.25, 0.53 and 0.24 probability of having weak class of degree, average class of degree and good class of degree respectively. In this case, monitoring is required to minimize poor performance either to sustain him/her in the average students cluster or improvement to good students cluster.

6. Clustering of At-Risks Students

The students' degree of membership in the weak students' cluster depicted in Table 6, was the vector used in determining the natural number of clusters, and in the actual clustering of at-risk students via the k-means algorithm. The Silhouette plot presented in Figure 8 shows that the optimal number of clusters for the dataset is four (4).

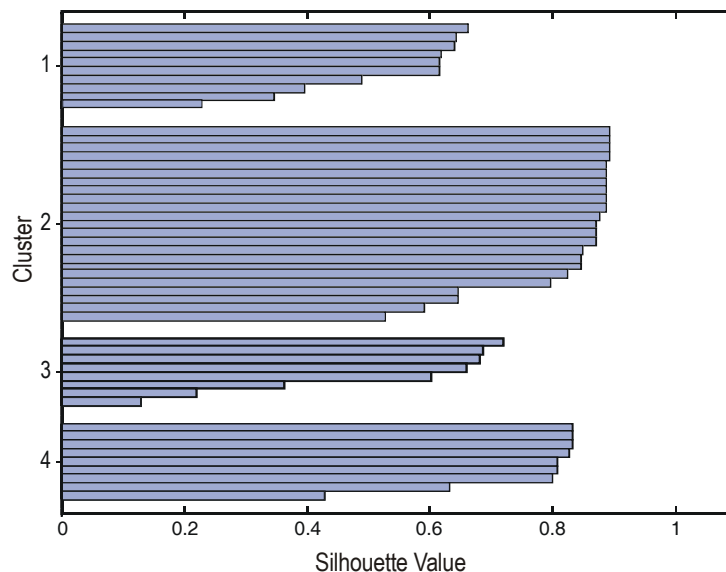


Figure 8. Silhouette plot for at-risk students clusters

As shown in Figure 8, most points in all the clusters have a large silhouette value, greater than 0.61, indicating that the clusters are well separated from each other. In addition, the mean silhouette value of 0.70 also proves that the number of clusters is optimal. The result of the classification of at-risk students into clusters with their silhouette values is presented in Table 7. As shown in Table 7, out of the 50 students in the test dataset, 23 students are in cluster 1 with no risks of performing poorly or having a weak class of degree. These students have very high chances of graduating with a minimum of second class lower; therefore monitoring of this category of students is required to sustain them in their respective clusters. The second cluster comprises student whose degree of membership in the weak students cluster is in the range [0.23, 0.51]. These students are those who have competing degree of membership in Weak and Average students clusters, hence require intervention plans to reduce their chances of belonging to the weak students cluster. The linguistic value indicating at-risks level of this cluster is *slightly risky*. In cluster 3, students have high probability of poor performance and are likely to spend more than the stipulated minimum duration of the programme; these students should be closely monitored for possible improvement to reduce their chances of retrogressing into the fourth cluster.

Table 7. At-Risk level clustering of students

S/N	Student ID	*DMWSC	Cluster Number	Linguistic Value of Risk Level	Silhouette Value	S/N	Student ID	*DMWSC	Cluster Number	Linguistic Value of Risk Level	Silhouette Value
1	11	0	1	Not Risky	0.85	26	45	0.33	2	Slightly Risky	0.66
2	10	0.01			0.87	27	43	0.34			0.69
3	25	0.01			0.87	28	36	0.38			0.72
4	41	0.01			0.87	29	17	0.43			0.69
5	2	0.02			0.89	30	34	0.46			0.61
6	6	0.02			0.89	31	47	0.51			0.36
7	18	0.02			0.89	32	22	0.62	3	Risky	0.35
8	32	0.02			0.89	33	38	0.7			0.62
9	9	0.03			0.89	34	46	0.72			0.65
10	24	0.03			0.89	35	33	0.76			0.66
11	26	0.03			0.89	36	1	0.79			0.64
12	28	0.03			0.89	37	21	0.8			0.62
13	29	0.04			0.89	38	23	0.8			0.62
14	31	0.04			0.89	39	20	0.82			0.49
15	8	0.05			0.88	40	50	0.83			0.40
16	30	0.07			0.85	41	7	0.86			0.09
17	39	0.07			0.85	42	3	0.89	4	Very Risky	0.43
18	15	0.08			0.83	43	49	0.91			0.63
19	12	0.09			0.80	44	4	0.94			0.81
20	35	0.13			0.65	45	14	0.94			0.81
21	40	0.13			0.65	46	19	0.95			0.83
22	13	0.14			0.59	47	27	0.98			0.83
23	44	0.15			0.53	48	42	0.98			0.83
24	37	0.23	2	Slightly Risky	0.13	49	48	0.98			0.83
25	16	0.24			0.22	50	5	0.99			0.81

* Degree of Membership of Students in the Weak Students Cluster.

The at-risk level of cluster 4 is *Very Risky*, with likelihood of having a pass class of degree above 88%. This category of students is at risk of attrition and may be advised to change or withdraw from the programme. A similar approach described in (Ekong, Inyang, & Onibere, 2012) guided the choice of triangular membership function. The fuzzy membership function for the at-risks students clusters, is presented in Equation 6.

$$\mu(x) = \begin{cases} \text{'Not Risky'} & \text{if } 0 \leq x \leq 0.15 \\ \text{'Slightly Risky'} & \text{if } 0.15 \leq x \leq 0.51 \\ \text{'Risky'} & \text{if } 0.51 < x \leq 0.86 \\ \text{'Very Risky'} & \text{if } 0.86 < x \leq 1.0 \end{cases} \quad (6)$$

7. Conclusion

Predicting students' performance is useful in identifying students who are likely to perform poorly in their studies. Fuzzy clustering technique has been used to perform important analysis in the educational environment for decisions to enhance educational standards. Significant first year courses based on their percentage contribution to graduating class of degree were identified and used for the experiment. The proposed system provides a grouping of students based on their level of performance and discovers the degree of membership of students in each cluster. Four (4) natural clusters of at-risk students were automatically identified with k-means algorithm. Results show that Sugeno-type inference system is best suitable for the provision of initial parameters for ANFIS training of students' dataset. In addition, FCM and ANFIS models provide an intelligent way of predicting and classifying students' academic performance. The results also prove the effectiveness of the combination of FCM and k-means algorithm and ANFIS in the classification of students based on academic performance and at-risk levels. As further work, the fuzzy membership functions and ANFIS rules could be used in the knowledge base of expert systems in the domain of students performance estimation and at-risk level prediction. These applications will help instructors, educational managers and students to improve the quality of performances by discovering at-risk students, at an early stage of their academic career and then develop a plan for minimizing poor performance and maximizing opportunities for excellent performance.

References

- Aali, K. A. (2009). Estimation of Saturation Percentage of Soil Using Multiple Regression, ANN, and ANFIS Techniques. *Computer and Information Science*, 2(3), 127-136.
- Baboo, S. S., & Priya, P. S. (2013). Clustering based integration of personal information using Weighted Fuzzy Local Information C-Means Algorithm. *International Journal of Advanced Trends in Computer Science and Engineering*, 2(2), 11-14.
- Bisht, D. C. S., & Jangid, A. (2011). Discharge Modeling using Adaptive Neuro - Fuzzy Inference System. *International Journal of Advanced Science and Technology*, 31(2011), 99-114.
- Castro, F., Nebot, A., & Mugica, F. (2007). Extraction of Logical Rules to Describe Students Learning Behaviour. *Proceeding of the sixth IASTED International Conference Web Based Education. Chamonix, France* (pp. 164-169).
- Chamillard, A. T. (2006). Using Student Performance Predictions in a Computer Science Curriculum. ITiCSE Bologna Italy. ACM (pp. 1-5).
- Chan, A. Y., Chow, K. O., & Cheung, K. S. (2008). Online Course Refinement through Association Rule Mining. *Journal of Educational Technology Systems*, 36(4), 433-444.
- Chandra, E., & Nandhini, K. (2010). Knowledge Mining from Student Data. *European Journal of Scientific Research*, 47(1), 156-163.
- Ekong, V. E., Inyang, U. G., & Onibere, E. A. (2012). Intelligent Decision Support System for Depression Diagnosis Based on Neuro-fuzzy-CBR Hybrid. *Modern Applied Science*, 6(7), 79-88. <http://dx.doi.org/10.5539/mas.v6n7p79>
- Ekong, V. E., Onibere, E. A., & Imianvan, A. A. (2011). Fuzzy Cluster Means System for the Diagnosis of Liver Diseases. *International Journal of Computer Science and Technology*, 2(3), 205-209.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*.
- Giovanis, E. (2010). A Study of Panel Logit Model and Adaptive Neuro-Fuzzy Inference System in the Prediction of Financial Distress Periods. *World Academy of Science, Engineering and Technology*, 40(2010), 651-657.
- Ghosh, S., & Dubey, S. K. (2013). Comparative Analysis of K-Means and Fuzzy C Means Algorithms. *International Journal of Advanced Computer Science and Applications*, 4(4) 34-39.
- Han, J. M., Kamber, & Tung, A. K. H. (2001). Spatial Clustering Methods. In H. Miller & J. Han (Eds.), *Data Mining: A Survey in Geographic Data Mining and Knowledge Discovery*. Taylor and Francis.
- Hossain, S. J., & Ahmad, N. (2012). Adaptive Neuro-Fuzzy Inference System (ANFIS) Based Surface Roughness Prediction Model for Ball End Milling Operation. *Journal of Mechanical Engineering Research*, 4(3), 112-129. <http://dx.doi.org/10.5897/JMER10.079>
- Imianvan, A. A., & Obi, J. (2012). Fuzzy Cluster Means Expert System for the Diagnosis of Tuberculosis.

Global Journal of Computer Science and Technology, 11(6), 43-51.

- Inyang, U. G. (2011). Generating Rules for Students' Performance Prediction in Tertiary Institutions. *ICASTOR Journal of Mathematical Sciences*, 5(2), 205-216.
- Inyang, U. G., Njumbwen, E., & Inyang, M. U. (2009). Design of An Analytic Hierarchy Process Based Decision Support System for Residential Property Renting. *Indian Centre for Advanced Scientific and Technology Research (ICASTOR) Journal of Mathematical Sciences*, 3(2), 183-195.
- Inyang, U. G., & Akinyokun, O. C. (2006). Fuzzy Cluster Means System for Matching Human Behavior Patterns to Ethnic groups. *Proceedings of the International Conference on New Trends in Mathematics and Computer Science with Applications to Real World Problems (NTMCS2006)*, Covenant University, Ota, Nigeria (pp. 429-434).
- Kifaya, S. Q. (2009). Mining Student Evolution Using Associative Classification and Clustering. *Communications of the IBIMA*, 11(2), 19-24.
- Kumar, P., Verma, P., & Shorma, R. (2010). Comparative Analysis of Fuzzy C Mean and Hard C Means Algorithm. *International Journal of Information Technology and Knowledge Management*, 2(2010), 1-5.
- Kumar, V., & Chadha, A. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. *International Journal of Advanced Computer Science and Applications*, 2(3), 80-84.
- Kumari, N., Sharma, B., & Gaur, D. (2012). Implementation of Possibilistic Fuzzy C-Means Clustering Algorithm in Matlab. *International Journal of Scientific & Engineering Research*, 3(11), 1-9.
- Mahdavi, Z., & Khademi, M. (2012). Prediction of Oil Production with: Data Mining, Neuro-Fuzzy and Linear Regression. *International Journal of Computer Theory and Engineering*, 4(3), 446-447. <http://dx.doi.org/10.7763/IJCTE.2012.V4.504>
- Miller, H. J., & Han, J. (2001). Geographic Data Mining and Knowledge Discovery-An Overview. In H. J. Miller & J. Han (Eds.), *Geographic Data Mining and Knowledge Discovery* (pp. 3-32). Taylor and Francis. http://dx.doi.org/10.4324/9780203468029_chapter_1
- Muzzammil, M. (2010). ANFIS Approach to the Scour Depth Prediction at a Bridge Abutment. *Journal of Hydroinformatics*, 12(4), 474-485. <http://dx.doi.org/10.2166/hydro.2010.004>
- Ngor, E. N. (2007). Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. *Fourth Congress of Electronics, Robotics and Automotive Mechanics. IEEE Computer Society* (pp. 354-359).
- Omolehin, J. O., Oyelade, J. O., Ojeniyi O. O., & Rauf, K. (2005). Application of Fuzzy Logic in Decision Making on Students' Academic Performance. *Bulletin of Pure and Applied Sciences*, 24(2), 281-187.
- 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.
- Paris, I. H. M., Affendey, S. L., & Mustapha, N. (2010). Improving Academic Performance Prediction using Voting Technique in Data Mining. *World Academy of Science, Engineering and Technology*, 62, 820-823.
- Petratis, P. S., Dunham, A. E., & Niewiarowski, P. H. (1996). Inferring Multiple Causality: the Limitations of Path analysis. *Functional Ecology*, 10, 421-431. <http://dx.doi.org/10.2307/2389934>
- Piatetsky-Shapiro, G. (2007). Data Mining and Knowledge Discovery 1996 to 2005: Overcoming the Hype and Moving from "University" to "Business" and "Analytics". *Data Mining and Knowledge Discovery*, 15(1), 99-105. <http://dx.doi.org/10.1007/s10618-006-0058-2>
- Rao, V. S., & Vidyavathi, S. (2010). Comparative Investigations and Performance Analysis of FCM and MFPCM Algorithms on Iris data. *Indian Journal of Computer Science and Engineering*, 1(2), 145-151.
- Romero, C., & Ventura, S. (2007). Educational Datamining: A Survey from 1995-2005. *Expert Systems with Applications*, 33, 135-146. <http://dx.doi.org/10.1016/j.eswa.2006.04.005>
- Romero, C., Ventura, S., & Garcia, E. (2008). Datamining in Course Management Systems: Moodle Case Study and Tutorial. *Computers & Education*, 51(1), 368-384. <http://dx.doi.org/10.1016/j.compedu.2007.05.016>
- Saltelli, A. (2002). Sensitivity Analysis for Importance Assessment. *Risk Analysis*, 22(3), 579-590. <http://dx.doi.org/10.1111/0272-4332.00040>

- Sansgiry, S. S., Bhosle, M., & Sail, K. (2006). Factors that Affect Academic Performance Among Pharmacy Students. *American Journal of Pharmaceutical Education*, 231-243.
- 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.
- 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.
- Yang, M. S. (1993). A Survey of Fuzzy Clustering. *Mathematical and Computer Modeling*, 18(11), 1-16. [http://dx.doi.org/10.1016/0895-7177\(93\)90202-A](http://dx.doi.org/10.1016/0895-7177(93)90202-A)
- Yayar, R., Hekim, M., Yilmaz, V., & Bakirci, F. (2011). A comparison of ANFIS and ARIMA Techniques in the Forecasting of Electric Energy Consumption of Tokat Province in Turkey. *Journal of Economic and Social Studies*, 1(2), 87-112.
- Zuviria, M. N., & Deepa, M. (2013). A Robust Fuzzy Neighborhood Based C Means Algorithm for Image Clustering. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(3), 87-94.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).