Soft Clustering Technique on Academics Performance Evaluation

**Jelili Oyelade[1,*+], Oladipupo O. O[1+,] Itunuoluwa Isewon[1]
& Obagbuwa[2] I. C.**

[1]Department of Computer and Information Sciences
Covenant University, Ota, Nigeria.

[2]Department of Computer Sciences
Lagos State University, Lagos, Nigeria

*** *Corresponding Author; ola.oyelade@covenantuniversity.edu.ng.*
+ *Joint first author*

Abstract: Clustering techniques are unsupervised learning methods of mining complex and multi-dimensional data sets such that observations in the same cluster are similar in some sense. The student academic performance evaluation problem can be considered as a clustering problem where clusters are formed on the basis of students intelligence. Choosing the right clustering technique for a given dataset is a research challenge. Therefore, intelligence-based grouping is essential for maintaining the homogeneity of the group; otherwise it would be difficult to provide good educational recommendation to the highly diverse student population. Homogenous grouping of students with similar result ranking into classes would further make student academic performance analysis detailed and sufficient for recommendation. Grouping of students using Fuzzy C-Means (FCM) techniques with the level of their degree of membership into different clusters allows for overlapping of boundaries and resolve sharp boundary problems as opposed to crisp-based method. FCM technique will reveal the degree of membership trend in the clusters which is the focus of this work. In this work, we implemented Soft clustering technique (Fuzzy C-Means) in C++ for student academic performance analysis. This will proffer recommendations that will enhance student performance.

Key words: K-means, Fuzzy-C- mean, Clustering algorithm, Performance evaluation

1.0 Introduction

Academic performance (AP) is the outcome of education, that is, the extent to which students has achieved in their educational goals. Academic performance have been linked to differences in intelligence. Students with higher mental ability tend to achieve highly in academic settings. AP has become the gatekeeper to institutions of higher education, shaping career paths and individual life trajectories(Stumm *et al.*, 2011). Student's academic performance is affected by numerous factor such as gender, age, teaching faculty etc. Many researchers conducted detailed studies about factors contributing the student performance at different levels. According to Minnesota (2007), the higher education performance is depending upon the academic performance of graduate students. Staffolani and Bratti (2002) observed that the measurement of students previous educational outcomes are the most important indicators of students future performance which implies that as the higher previous appearance, the better the student's academic performance in future endeavours.

Students enrolled for a course in an institution have to complete the minimum number of courses required before graduating. These courses are only completed if they meet all requirements and pass with an acceptable grade. A student that fails a course earns no credit for that course.

The academic performance of a student is based on their GPA which is the average number of points the student attains in all their courses graded from A –F and this in turn determines the overall success of the student in their program of study.

Student academic performance can be seen as a clustering problem where each cluster is represented based on the intelligence of the student. This is needed especially in a diverse student population to ensure uniformity. This uniform grouping would make results more feasible and a basis for comparison can also be established. Using this clustering technique, the areas of strength and weakness of the student can be revealed so that proper monitoring can be established.

Grouping or clustering students using fuzzy-based techniques with the level of their degree of membership into different clusters may be a realistic approach as opposed to crisp-based methods (e.g. k-means). For example, a student with scores 30, 50, 60, and 70 will be in the region of good performance using k-means approach in Oyelade, *et al* (2010); but this FCM technique will reveal the degree of membership trend in the clusters which may not necessary be in good performance state.

2.0 Literature Review

Partitioning methods aim to find the best partition of data into k clusters in such a way that one criterion is optimized. The research work by Anand *et. al.*, (2009) only provides Data Mining framework for Students' academic performance. The research by Varapron et al, (2003) used rough Set theory as a classification approach to analyze student data where the Rosetta toolkit was used to evaluate the student data to describe different dependencies between the attributes and the student status where the discovered patterns are explained in plain English. Oyelade, *et. al.*, (2010) applied k-means technique with deterministic approach to student's academic performance into different k clusters but fail to reveal each student's area of strength and weakness in different clusters with respect to the degree of membership to each cluster. Ramjeet and Ahmed (2012) proposed a dynamic fuzzy expert system to analyze and find modelling academic performance to improve on the quality of students and teachers performance in academic domain but failed to reveal the degree of membership strength in difference clusters. In SajadinSembiring (2011), they applied Smooth Support Vector Machine (SSVM) classification and kernel k-means clustering algorithms by employing psychometric factors as variables predictors where their results showed a model of student academic performance predictors. Sharma (2013) presented a data mining techniques to process a dataset and identify the relevance of classification on the test data. Durairaj and Vijitha, (2014) used WEKA tool for prediction of student's performance in term of pass percentage and fail percentage using K-Means clustering algorithm.

In this work, we implemented fuzzy c-mean algorithm (Bezdek, 1981) in C++ for partitioning of students academic results with the level of their degree of membership into different clusters . In addition to the specification of the number of k clusters in the data set, the FCM method requires to choose m, which is the fuzziness parameter. There is little literature on the choice of this parameter (Bezdek, 1981; McBratney and Moore, 1985) but this is not the focus of this work.

3. Materials and Methods

We demonstrated our technique on student's result data set with nine courses offered in a semester from a private university in Nigeria. The total number of 79 students were considered and analyzed using FCM algorithm.

3.1 Development of FCM algorithm

The crisp clustering methods assign each object to one cluster only, unlike fuzzy clustering methods, it assigned each object to one or more cluster depending on the degree of membership in that cluster. The degree of membership has values ranging from 0 to 1. If the degree of membership of an object in a particular cluster is very close to 1, this indicate a very strong association of an object in that cluster and values close to 0 indicate weak or absent association with the cluster. The fuzzy c-means algorithm (FCM) (Bezdek, 1981) is one of the most widely used methods in fuzzy clustering which is based on the concept of fuzzy c-partition, introduced by Ruspini (1969) as follows.

Assume a set of n objects $X = \{x_1, x_2, \ldots, x_n\}$, where x_i is a d-dimensional point. A fuzzy clustering is a collection of k clusters C_1, C_2, ..., C_k and a partion matrix $U_{i,j} = u_{i,j} \in [0,1]$ for i = 1, ..., n and j = 1, ..., k, where each element $u_{i,j}$ is a weight that represents the degree of membership of object i in cluster C_j., all weight for a given point x_i must add up to 1. That is, $C_j = \dfrac{\sum_{i=1}^{n} u^{m}{}_{ij} x_i}{\sum_{i=1}^{n} u^{m}{}_{ij}}$ such that each cluster C_j contains non-zero weight, i.e. $0 < \sum_{i=1}^{n} u_{i,j} < n$.

Like k-means, FCM also attempts to minimize the sum of the squared error (SSE). That is,

In k-means:

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_i} dist(C_i, x)^2$$

In FCM:

$$U_{ij} = \frac{(1/dist(x_i, C_j)^2)^{1/(m-1)}}{\sum_{q=1}^{k}((1/dist(x_i, C_q)^2)^{1/(m-1)})}$$

where m is the parameter that determines the influence of the weights and $m \in [1, \dots, \infty]$.

For a cluster C_j, the corresponding centroid C_j is calculated as follows:

$$C_j = \frac{\sum_{i=1}^{n} u^m{}_{ij} x_i}{\sum_{i=1}^{n} u^m{}_{ij}}$$

This is an extension of the centroid in k-means. The difference here is that all points are considered and the contribution of each point to the centroid is weighted by its membership degree.

The fuzzy partition update can be obtained by minimizing the SSE subject to the constraint that the weights sum to 1. That is:

$$U_{ij} = \frac{(1/dist(x_i, C_j)^2)^{1/(m-1)}}{\sum_{q=1}^{k}((1/dist(x_i, C_q)^2)^{1/(m-1)})}$$

U_{ij} should be high if x_i is close to the centroid C_j, i.e. if $dist(x_i, C_j)$ is low.

The effect of parameter m in FCM is stated as follows:

- If m>2, then the exponent 1/(m-1) decrease the weight assigned to clusters that are close to the point.
- If $m \to \infty$, then the exponent $\to$0. This implies that the weights $\to$1/k.
- If m $\to$1, the exponent increases the membership weights of points to which the cluster is close. As m $\to$1, membership $\to$0, for all the other clusters.

3.1.1 The algorithm steps

Given a dataset of n data points $X = \{x_1, x_2, \dots, x_n\}$ such that each data point is in $\boldsymbol{R}^n$, the problem of finding the minimum J_m is given as:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} U_{ij}^m \left\| x_i - c_j \right\|^2 \quad 1 \leq m \leq \infty \qquad (1)$$

- m is the fuzziness parameter which regulate the degree of membership in the clustering process; for $m = 1$, the problem is the classical minimum sum of squares clustering and the partition is crisp. Therefore, m is any real number > 1;
- $U_{i,j}$ is the degree of membership of x_i in the cluster j;
- x_i is the $i - th$ the dimensional measured data.

- c_j is the dimensional center of cluster

Therefore, fuzzy partioning is carried out through iterative optimization of the objective function J_m depicted in equation 1 above, with the update membership $U_{i,j}$ and the cluster centers c_j described by:

$$U_{i,j} = \frac{1}{\sum_{j=1}^{c}\left(\frac{||x_i - c_j||}{||x_i - c_k||}\right)^{\frac{2}{m-1}}}$$ and

$$c_j = \frac{\sum_{i=1}^{N}(U_{i,j}^{m} x_i)}{\sum_{i=1}^{N} U_{i,j}^{m}}$$

This iteration will stop when:

$max_{i,j} = \{|U_{i,j}(k+1) - U_{i,j}(k)|\} \leq \epsilon$ where ϵ is the termination criterion between 0 and 1 and k is the iteration steps.

The algorithm steps is described as follows:

1. Initialize $U = [U_{i,j}]$ matrix. i.e. $U^{(0)}$
2. At k-steps:
   a. Calculate vector $C^k = [c_j]$ with U^k i.e.
   $$c_j = \frac{\sum_{i=1}^{N}(U_{i,j}^{m} x_i)}{\sum_{i=1}^{N} U_{i,j}^{m}}$$
3. Update:

$$U_{i,j} = \frac{1}{\sum_{j=1}^{c}\left(\frac{||x_i - c_j||}{||x_i - c_k||}\right)^{\frac{2}{m-1}}}$$

4. If $||U_{i,j}(k+1) - U_{i,j}(k)|| \leq \epsilon$ stop else return to step 2.

4. Results and Disscussion

From the fuzzy C means analysis we have 4 clusters (cluster 0 to 3) from the academic performance point of view each cluster representation is shown in Table 1

Table 1: Fuzzy Clusters academic performance representation

Cluster number	Grade performance	Linguistic performance	Class of Honour
Cluster 0	A & B	Good	2nd class upper and above
Cluster 1	F	Poor	Fail
Cluster 2	C	Average	2nd class lower
Cluster 3	D	Fair	3rd class

This can be represented in a fuzzy linguistic model such that the linguistic variable is student performance and the fuzzy sets are Good, Poor, Average and Fair:

Student Performance {Good, Poor, Average, Fair}

A sample data of 76 records with 9 attributes was used. Each record represents an instance of a student percentage quantitative performance in 9 core courses offered in a particular session. With fuzzy -c means analysis the system was able to cluster each student in their best performance cluster. Also, it reveals each record membership function in each cluster. The system assigns membership value to each data point (each record) corresponding to each cluster center on the basis of distance between the cluster and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. The summation of membership of each data point should be equal to one. This reveals each student strength distribution across the 4 categories of performances. For instance Table 2 shows an instance of the fuzzy-cC means analysis result. The percentage strength distribution for each data point in each cluster is shown in Table 3. Figure 1 shows a graphical distribution of the student's strength.

Table 2: An instance of Fuzzy C means student performance analysis result with Data point cluster.

Record Number	Cluster 0 (Good)	Cluster 1 (Poor)	Cluster 2 (Average)	Cluster 3 (Fair)	**Record Cluster**
Data [2]	0.55	0.01	0.23	0.21	Cluster 0
Data [9]	0.79	0.01	0.10	0.10	Cluster 0
Data [57]	0.09	0.01	0.44	0.46	Cluster 3
Data [43]	0.23	0.02	0.38	0.37	Cluster 2
Data [73]	0.03	0.87	0.05	0.05	Cluster 1

Table 3: An instance of Fuzzy C means student performance analysis percentage strength distribution

Record Number	Cluster 0 (Good)	Cluster 1 (Poor)	Cluster 2 (Average)	Cluster 3 (Fair)
Data [2]	55%	1%	23%	21%
Data [9]	79%	1%	10%	10%
Data [57]	9%	1%	44%	46%
Data [43]	23%	2%	38%	37%
Data [73]	3%	87%	5%	5%

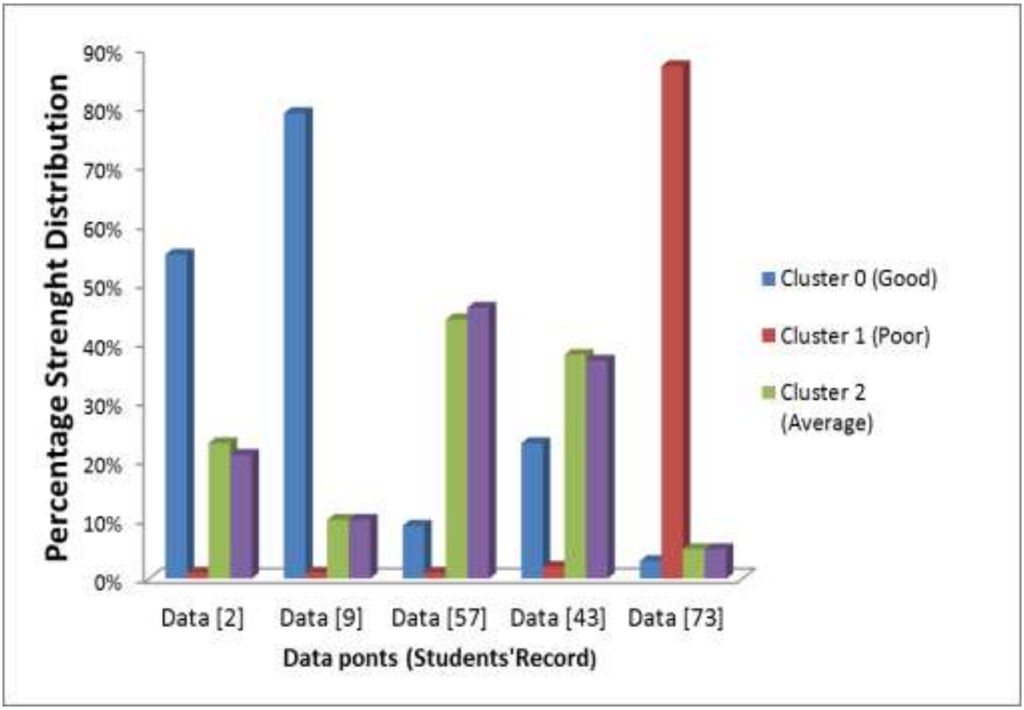

Figure 1: Graphical representation of Student's strength distribution

From the Table 2 Data[2] student has 55% of his strength in Cluster 0 which represents grade B-A ; good performance, 10 % in Cluster 1 which represents grade F; poor performance , 23% in cluster 2 of grade C; Average performance and 21% in cluster 3 which represents grade D; fair performance. These strength distributions show that this student is not a stable good student. He needs to improve his study capacity so as to strengthen his good performance ability; hence he may fall into average or below average performance category.

Data [9] student has 79% of his strength in cluster 0, of good performance, 1% in poor performance, 10% in average and fair performances. These show that the student is a stable good student. He just needs to maintain his performance. It might be difficult for him to move below average.

Data [59] student has 9% of his strength in good performance, 1% in poor, 44% in average and 46 % in fair performance. These show that this student is below average. Though he might not graduate with second class upper and above but if he works harder he can still increase his chances of second class lower, hence he might end up as a fair student of 3rd class.

Data [43] student has 23% of his strength in good, 2% in poor, 38% in average and 37% in fair. This reveals that this kind of student is above average but due to his carelessness his performance is fair. With proper monitoring and advice he can increase his good performance ability while he moves away from fair performance. Nevertheless, if care is not taking he might graduate with 3rd class.

Data [73] has 3% of his strength in good, 87% in poor, 5 % in average and fair performances. It is obvious that this student needs not to be promoted if after a session he has these strength distributions. He must have gathered enough carryovers, then he needs to be advised on time to change his course or withdraw without wasting time and resources.

Finally, the overall performance of this set of students is represented in Figure 2. It can be concluded that most of the students in this set are stable good students. Few of them, precisely 4 are poor students. Also, the graph reveals that those students that fall into average and fair performances have a thin line difference. The implication is that, if a student is in average performance this session under consideration, if care and diligent effort is not invested in the following sessions the student can easily fall into 3rd class category and also, if a student is in 3rd class, with little diligent effort this student can move to second class lower.

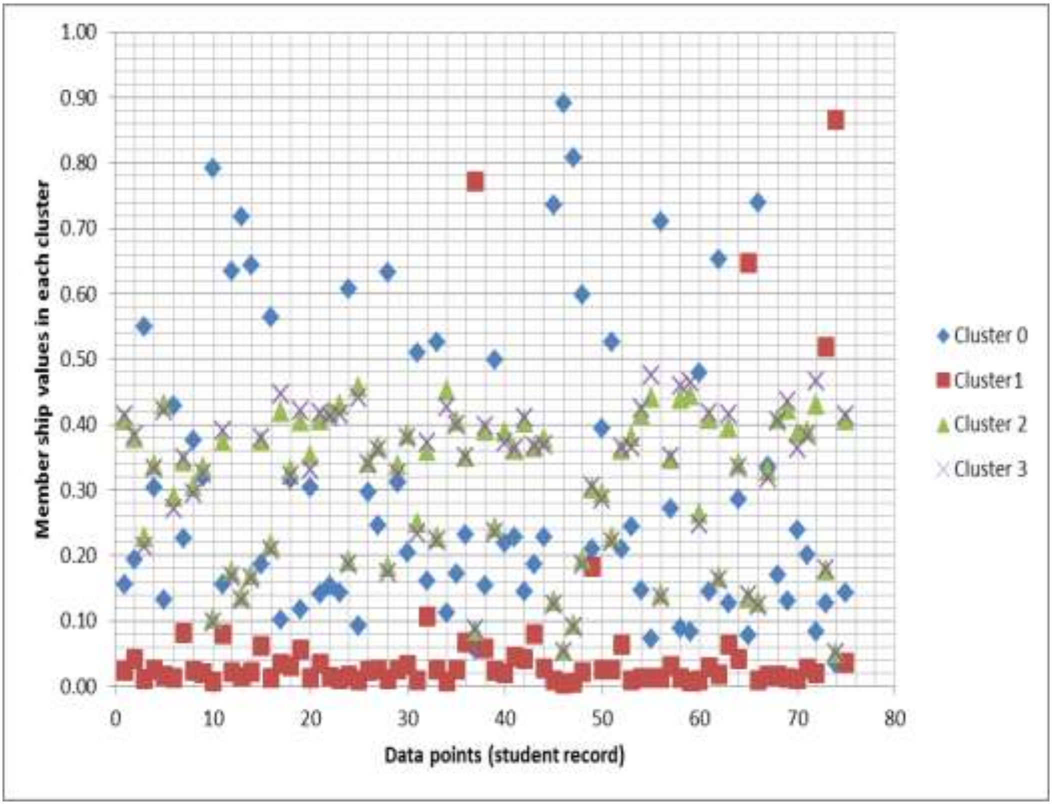

Figure 2: Student overall performance chart

5. Conclusion

In this paper, we implemented the qualitative power of FCM clustering algorithm in C++ to demonstrate the importance of degree of membership of student's performance in different clusters. This reveals each student's area of strength and weakness which the hard clustering technique (k-mean) fail to reveal (Oyelade, et. al., 2010). This model improved on some of the limitations of the existing methods. For example, the research work by Anand *et. al.*, (2009) only provides Data Mining framework for Students' academic performance. The research by Varapron *et al* (2003) used rough Set theory as a classification approach to analyze student data where the Rosetta toolkit was used to evaluate the student data to describe different dependencies between the attributes and the student status where the discovered patterns are explained in plain English.

Therefore, FCM clustering algorithm serve as a good benchmark in monitoring the progress of students' performance in the institutions which enhances the decision making by academic planners by monitoring the candidates' performance semester by semester by improving on the future academic results in the subsequence academic session.

6. References

Bezdek, J. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Ruspini, E. (1969). A new approach to clustering. Information and Control, 15, 22–32..

McBratney, A.B. and Moore, A.W. (1985). Application of fuzzy sets to climatic classification. Agricultural and Forest 24 Meteorology, 35, 165-185.

Ramjeet and Ahmed (2012): Academic Performance Evaluation Using Fuzzy C-Means. International Journal of Computer Science Engineering and Information Technology Research. 2(4), 55-84.

Sajadin Sembiring, (2011) "Prediction Of Student Academic Performance by an Application of Data Mining Techniques", *International Conference on Management and Artificial Intelligence* IPEDR, IACSIT Press, 6.

Durairaj and Vijitha, (2014), Educational Data mining for Prediction of Student Performance Using Clustering Algorithms. International Journal of Computer Science and Information Technologies, 5(4), 5987-5991.

Sharma, T. C., (2013). WEKA Approach for Comparative Study of Classification Algorithm, (IJARCCE*) International Journal of Advanced Research in Computer and Communication Engineering*, 2(4).

Oyelade, O. J., Oladipupo, O. O., and Obagbuwa, I. C. (2010): Application of k-Means Clustering algorithm for prediction of Students' Academic Performance. International Journal of Computer Science and Information Security, Vol. 7(1), Pg. 292-295.

Anand, N. V., Kumar and Uma, G. V. (2009). Improving Academic Performance of Students by Applying Data Mining Technique. European Journal of Scientific Research, 34(4).

Varapron P. et al.(2003). Using Rough Set theory for Automatic Data Analysis. 29th Congress on Science and Technology of Thailand.

Von Stumm, Sophie; Hell, Benedikt; Chamorro-Premuzic, Tomas (2011): "The Hungry Mind: Intellectual Curiosity is the Third pillar of Academic performance", perspective on psychological Science, 6(6), 574-588.

Bratt M. and Staffolani, S. (2002). Student Time allocation and Educational production functions". University of Ancona, Department of Economics. Working paper no 170.

Minnesota Measures (2007). Report on higher education performance. Retrieved on May 24, 2008. www.opencongress.org/bill/110.s/642/show-139k.