International Journal of Communication and Information Technology

E-ISSN: 2707-6628
P-ISSN: 2707-661X
www.computersciencejournals.com/ijcit
IJCIT 2022; 3(1): 01-05
Received: 12-10-2021
Accepted: 03-12-2021

Keshinro Kazeem Kolawole
Department of Computer Engineering, Laspotech, Ikorodu, Lagos, Nigeria

Dr. Adenowo Adetokunbo O
Department of Electronics & Computer Engineering, LASU, OJO, Lagos, Nigeria

Sarumi Jamiu A
Department of Mechatronics Engineering, Laspotech, Ikorodu, Lagos, Nigeria

Balogun Wasiu Adebayo
Department of Mechatronics Engineering, LASPOTECH, Ikorodu, Lagos, Nigeria

Corresponding Author:
Keshinro Kazeem Kolawole
Department of Computer Engineering, Laspotech, Ikorodu, Lagos, Nigeria

Performance evaluation student result using k-means clustering

Keshinro Kazeem Kolawole, Dr. Adenowo Adetokunbo O, Sarumi Jamiu A and Balogun Wasiu Adebayo

DOI: https://doi.org/10.33545/2707661X.2022.v3.i1a.35

Abstract
The performance of students is a vital and significant element of institutions of higher learning, for both the student and the academic community as a whole. As a result, higher education institutions must be more flexible in terms of performance metrics and ideas. However, when the student population grows owing to sessional admission, obtaining accurate information on students' performance becomes more difficult due to the huge amount of data in educational databases (for about 1-100years). Clustering is one of the data mining techniques used to examine large amounts of data. It organizes data into clusters so that items are placed together in the same cluster if they are comparable based on certain criteria. Several methods for improving the performance of the K-means clustering algorithm used in big data analysis have been proposed in the literature, but the proposed modified K-means clustering algorithm is less time-consuming, more efficient, has less complexity, and, most importantly, produces better clustering. To categorize numerical data, the modified K mean method is employed. However, the data in each cluster may be susceptible to outliers and noisy data, which may decrease the accuracy rate of data matching, since pattern matching will not readily enable prediction of the cluster center and therefore cannot characterize the data in the cluster. The modified k-means clustering method, which is suitable for large data from social media, sensors, search engines, GPS, transaction/financial records, satellites, e-commerce sites, and other sources, is suggested to address the issue and assess the results produced.

Keywords: K-means clustering, academic community, performance metrics and ideas

Introduction

The ability to monitor the academic performance of students is a critical issue in the academic realm of higher learning. It is established a framework for assessing students' achievements based on cluster analysis, which employs standard statistical methods to arrange their score data according to the degree of their production. In this paper, we used the k-mean clustering method to analyze the results of students. The model was combined with the deterministic model to evaluate the students' outcomes of a private institution in Nigeria, which is a successful benchmark for tracking the advancement of academic success of students in higher institutions for academic planners to make an appropriate decision. (2014, "Modeling Academic Performance Evaluation Using Hybrid Fuzzy Clustering Techniques") Clustering methods divide the sample into several clusters (groups, sub-sets, and categories). Although no uniform definition exists, many scientists define a cluster in terms of internal homogeneity and exterior separation (Ghadiri et al., 2017) [3], which means that patterns inside the same cluster should be comparable, but patterns in other clusters should not be similar to each other (Yadav et al., 2014) [12]. As a result, the correct identification of clusters is dependent on how similarities are detected. Any distance function (more frequently, the dissimilarity) such as the Euclidean distance or the Mahalanobis distance is a generic characteristic of similarity. The choice of the measure of (dis)similarity causes the cluster to form and, as a result, determines the performance of a grouping algorithm in the application area. The community of clustering algorithms, for example, decides on hyper-sphere shaped or hyper-ellipsoidal clusters based on Euclidean and Mahalanobis distances. Normally, we are unaware of the most natural and effective methods. When we use a clustering method, we may cluster types or a particular dataset. - The dataset has a different data distribution than previous data sets, necessitating the use of various cluster types.

Related knowledge (patterns, records, etc.) for an academic environment are based on the environment's social, political, and economic circumstances.

However, the primary focus of a higher education institution is based on academic records such as attendance, internal mark assessment, seminar assessment, class assignment assessment, and the school's main examination. All carry some proportion of the overall marks, which are averaged to be 100 percent.

Experimentation was carried out on the existing and accessible database to validate the performance output in terms of accuracy, specificity, flexibility, and execution time (that is, the student data set). Performance assessment is a critical notion in developing predictive data that can be utilized to offer required facilities in the future. The primary purpose for planning for the future is to make decisions that will enhance happiness for the would-be person. Performance assessment of student outcomes enables us to comprehend and forecast not only the students but also the unit or institution from whence the results are coming. As a result, an enabling atmosphere for a creative future is created.

Several measures or measuring points are considered when reviewing the performance of an academic institution, such as the beautification of the institution, the availability of research materials in the institution, the state (social, political, and economic condition) of the academia, the number of journal publications by the academic staff, the number of laureates won by the academia, and so on. In all of these cases, the job description was never taken into account. As a result, student academic achievement that accurately reflects the institution's status should be prioritized.

In light of these considerations, the purpose of this study is to evaluate the academic performance of students in certain chosen departments in terms of accuracy, specialization, flexibility, and implementation time, to put the institution in the best possible position.

Materials and Method

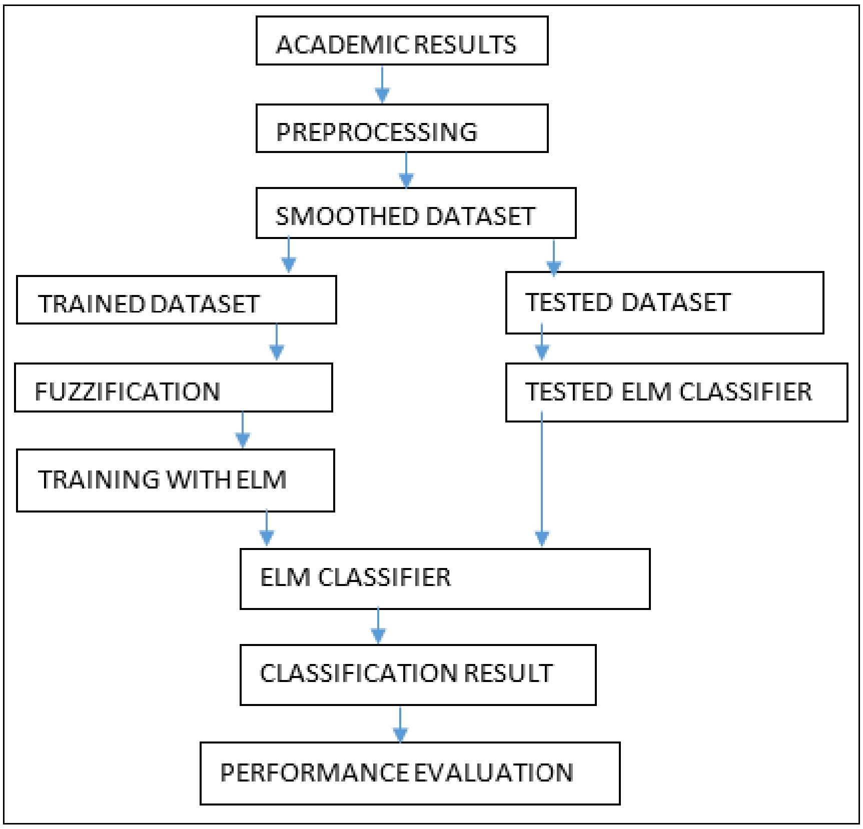


Fig 3.1: Flowchart for fuzzification of results

Overview

The dataset consisted of details of students of five consecutive years. The main features are the following attributes for each course attended by the student

i. Attendance
ii. Internal mark assessment
iii. Seminar assessment
iv. Class assignment assessment
v. Polytechnic marks scored

The dataset consisted of approximately 8000 records. The attributes internal assessment, seminar assessment, and the class assignment were transformed and consolidated into proper normal forms appropriate for mining. Normalization was done on these attributes so that data should fall within a small specified range and hence does not outweigh the measurement of other attributes.

The flow chart for the system is described above, and

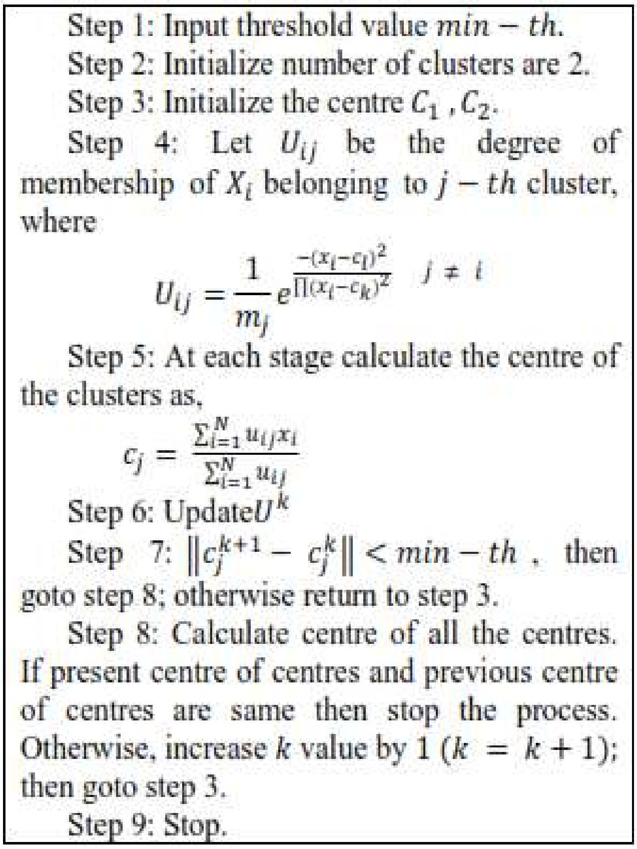
Proposed modified K-means clustering algorithm

The number of clusters cannot always be determined, however, or is feasible. The key goal of the FCM algorithm is to minimize the importance of an objective function to a minimum. The objective function tests the consistency of the separation of data into clusters. By measuring the distance from pattern location, the FCM algorithm tests the consistency Xi to the Main Cluster wj with template gap Xi to other centers of clusters. FCM is a system that requires the data entity to belong to two or more clusters. In the pattern recognition field, this approach is widely used. Dunn and Bezdeck 's approach, which is based on decreasing the objective function Eq. 1.

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} U_{ij}^{m} \,|\, X_i - c_j|^2, i \leq m \quad (1)$$

Where is seen to be more than 1. areal level. Bedeck was scheduled for 2.00, U is the membership standard of Xi in the company j, cj is at the core of $j - th$ cluster, and $\|\ast\|$ It's a 'width' norm. Fuzzy clustering is an iterative method of optimization, in which Uij and the cj cluster centers play an objective role in each iteration defined in Eq. 3.15.

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

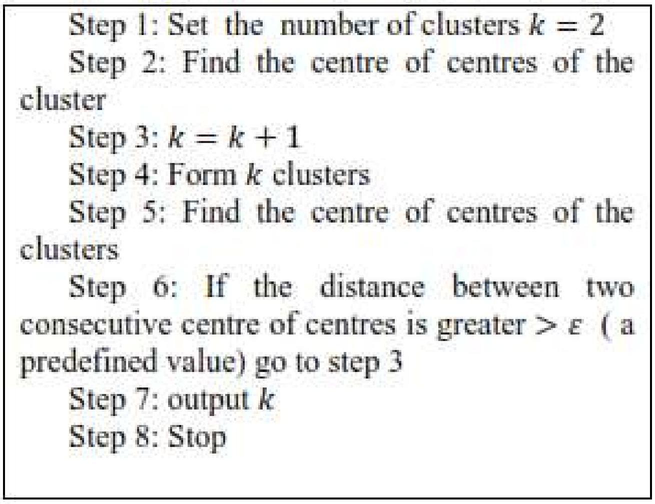


```
    Step 1: Input threshold value $min - th$.
    Step 2: Initialize number of clusters are 2.
    Step 3: Initialize the centre $C_1$ , $C_2$.
    Step 4: Let $U_{ij}$ be the degree of
membership of $X_i$ belonging to $j - th$ cluster,
where
        $U_{ij} = \frac{1}{m_j} e^{\frac{-(x_i - c_i)^2}{\prod(x_i - c_k)^2}} \quad j \neq i$
    Step 5: At each stage calculate the centre of
the clusters as,
        $c_j = \frac{\sum_{i=1}^{N} u_{ij} x_i}{\sum_{i=1}^{N} u_{ij}}$
    Step 6: Update$U^k$
    Step 7: $\|c_j^{k+1} - c_j^k\| < min - th$ , then
goto step 8; otherwise return to step 3.
    Step 8: Calculate centre of all the centres.
If present centre of centres and previous centre
of centres are same then stop the process.
Otherwise, increase $k$ value by 1 ($k = k + 1$);
then goto step 3.
    Step 9: Stop.
```

Fig 2: Algorithm for fuzzy clustering

And an optimization clustering process that recognizes clusters and assigns the objects to the nearest or related clusters is normally specified to minimize a certain calculation. The well-known centroid clustering partition method is one of the current methodologies of FCM and K-mean approaches. The artifacts are categorized based upon groups in a K-means strategy. The centroid or medium is the symbol of each cluster. If the data are real-life data, then the attribute vectors' arithmetic mean is optimal for all objects within a cluster. FCM is identical to k-means algorithms; on the other hand, initially. Proposed time and complexity saving algorithm that holds the middle pixel steady and iterates the adjacent pixels.

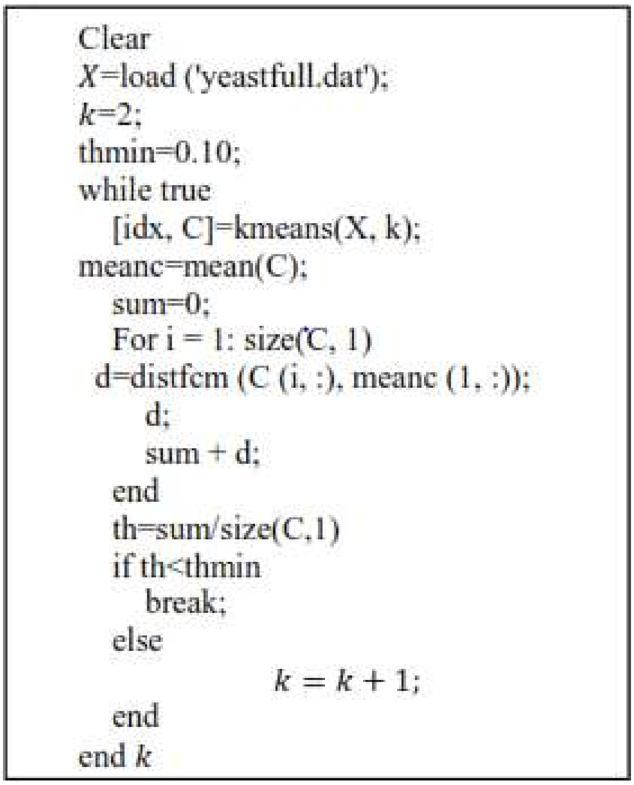
Below is a pseudo-code of the algorithm k-means modified as shown in Figure 3.



```
    Step 1: Set the number of clusters $k = 2$
    Step 2: Find the centre of centres of the
cluster
    Step 3: $k = k + 1$
    Step 4: Form $k$ clusters
    Step 5: Find the centre of centres of the
clusters
    Step 6: If the distance between two
consecutive centre of centres is greater $> \varepsilon$ ( a
predefined value) go to step 3
    Step 7: output $k$
    Step 8: Stop
```

Fig 3: Modified K-means Algorithm

The consistency of the method is further accomplished with the implementation of the equation (3.16) in the k-means algorithm. Below is the Matlab code for the algorithm k-means modified shown in figure 4.



```
Clear
X=load ('yeastfull.dat');
k=2;
thmin=0.10;
while true
   [idx, C]=kmeans(X, k);
meanc=mean(C);
   sum=0;
   For i = 1: size(C, 1)
  d=distfcm (C (i, :), meanc (1, :));
      d;
      sum + d;
   end
   th=sum/size(C,1)
   if th<thmin
      break;
   else
                  k = k + 1;
   end
end k
```

Fig 4: Matlab Code for Modified K-means

Result and Discussion

Experimental analysis of k-means and modified k-means

We applied the model on the data set (the academic result of one semester) of the computer engineering department, Lagos state polytechnic. The result generated is shown in tables 2, 3, and 4, respectively. In table 2, for k = 3; in cluster 1, the cluster size is 25 and the overall performance is 62.22. Also, the cluster sizes and the overall performances for cluster numbers 2 and 3 are 15, 29, and 45.73, and 53.03, respectfully. Similar analyses also hold for tables 3 and 4. The graphs are generated in figures 5, 6, and 7, respectively, where the overall performance is plotted against the cluster size.

Table 5 shows the size of the data set in the form N of M matrices, where N is the lines (number of the students) and M is the column (number of courses provided by each student. The overall performance is assessed by using a deterministic model equ (3)

$$\frac{1}{N}\left(\sum_{j=1}^{N}\left(\frac{1}{n-1}\sum_{i=1}^{n-1}X_i\right)\right) \qquad \text{equ (3)}$$

Where

N = the total number of students in a cluster and

n = the dimension of the data

Table 2: Performance index

70 and above	Excellent
60-69	Very Good
50-59	Good
45-49	Very Fair
40-45	Fair Below
Below 45	Poor

For cluster size 25, the overall performance of Figure 4.9 is 62.22%, while for the cluster size 15, the total performance of 45,73%, and for the cluster size 29, total performance was 53.03%. The results showed that 25 students out of 79 had a Very Good" (62.22 percent), while 15 out of 79 students had a very "Fair" performance (45.73 percent) and the 29 other students were "Good" (53.03%) as shown in table 2 of the index.

Figure 7 shows trends in the analysis of the performance as follows; for cluster size 24 overall performance is 50.08% while for group size 16 the total performance is 65.00%. The overall output of cluster size 30 is 58,89%, while cluster size 09 is 43,65%. The trends in this analysis showed that in the "good" index region in Table 2 above, there are 24 students (50.08 percent), while in the "very good" region there are 16 (65.00 percent). Thirty students performed well (58.89%) and nine students were performed fairly (43.65 percent).

For cluster size 19, the overall performance is 49,85%, whereas, for cluster size 17, the overall performance is 60,97%. The total performance of cluster size 9 is 43.65% whereas the total output of cluster size 14 is 64.93% and cluster size 20 is 55.79%. This performance analysis showed 19 students crossing the region of 'Good' (49.85%), while 17 had Very Good' performance results (60.97 percent). 9 students fall under the "Fair" Performance Index region (43.65%), 14 students are in the Very Good" (64.93%), and the other 20 have "Good Performance (55.79 percent).

Table 3: K = 3

Cluster #	Cluster size	Overall Performance
1	25	62.22
2	15	45.73
3	19	53.03

Table 3: K = 3

Cluster #	Cluster size	Overall Performance
1	24	50.08
2	16	65.00
3	30	58.89
4	9	43.65

Table 3: K = 5

Cluster #	Cluster size	Overall Performance
1	19	49.85
2	17	60.97
3	9	43.65
4	14	64.93
5	20	55.79

Table 3: K = 3

Student"s Scores	Number of Students	Dimension (Total number of courses)
Data	79	9

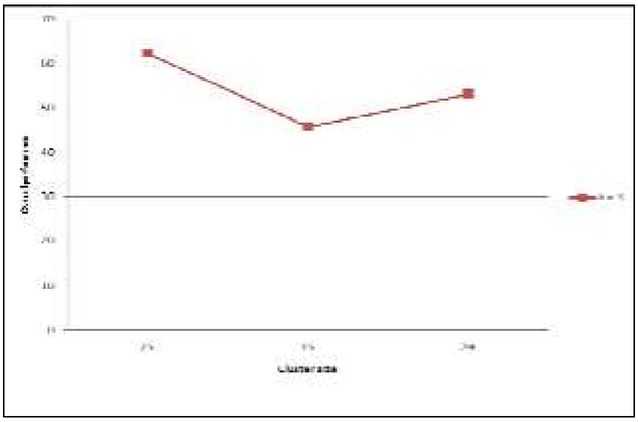


Fig 5: Overall Performance versus cluster size (# of students) k = 3

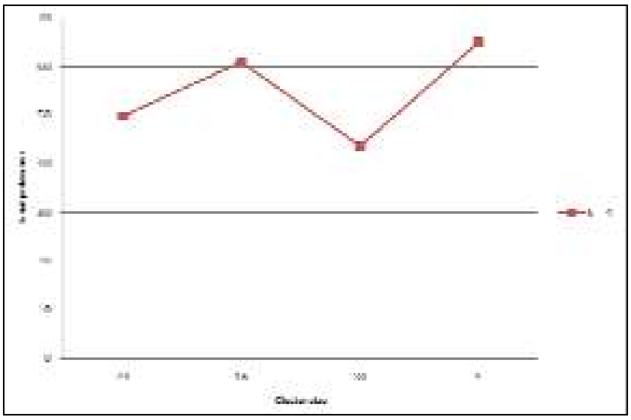


Fig 6: Overall Performance versus cluster size (# of students) k = 4

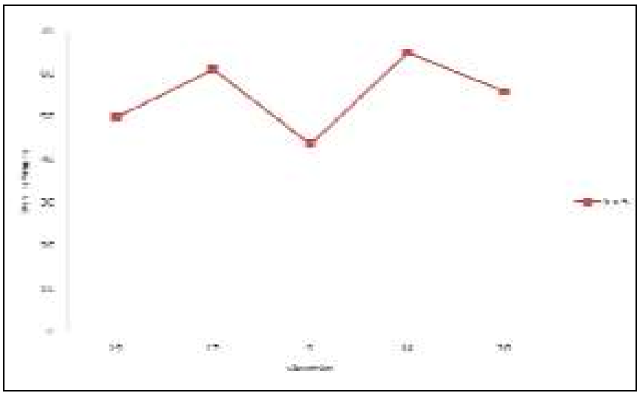


Fig 7: Overall Performance versus cluster size (# of students) k = 5

Conclusion

The clustering method to be used is entirely determined by the kind of data to be grouped and the aim of the clustering applications. A hard-clustering method, such as the K-

Means algorithm, is appropriate for a clustering job; while, a fuzzy clustering algorithm, such as FCM, is appropriate for overlapping clustering problems. In certain cases, we cannot assume that data belongs to just one cluster. It is conceivable that certain data characteristics contribute to several clusters. A document, as in the case of document clustering, may be classified into two groups. We usually choose membership value-based clustering, such as FCM, for these reasons.

In this article, we conducted a comparison study of the Fuzzy clustering algorithm and the K-Means method on behalf of the Hard-clustering algorithm.

Based on our tests, we discovered that the K-Means method takes less time to compute than the FCM algorithm for the Iris dataset. As a result, this study indicates that K-performance Mean's is superior to FCM's performance in terms of computing time. Because the fuzzy clustering method employs a greater number of fuzzy logic-based computations, its computational time rises in comparison.

References

1. Aruna Kumar SV, Harish BS, Mahanand BS, Sundararajan N. An efficient Meta-cognitive Fuzzy C-Means clustering approach. Applied Soft Computing Journal. 2019. https://doi.org/10.1016/j.asoc.2019.105838
2. Erik A, Kuvvetli Y. A new approach to supply chain performance assessment. Journal of the Faculty of Engineering and Architecture of Gazi University. 2020;35(4). https://doi.org/10.17341/gazimmfd.691906
3. Ghadiri N, Ghaffari M, Nikbakht MA. BigFCM: Fast, precise and scalable FCM on hadoop. Future Generation Computer Systems. 2017;77. https://doi.org/10.1016/j.future.2017.06.010
4. Jain N, Singh AR. Sustainable supplier selection under must-be criteria through Fuzzy inference system. Journal of Cleaner Production. 2020, 248. https://doi.org/10.1016/j.jclepro.2019.119275
5. Kesarwani A, Khilar PM. Development of trust based access control models using fuzzy logic in cloud computing. Journal of King Saud University - Computer and Information Sciences. 2019. https://doi.org/10.1016/j.jksuci.2019.11.001
6. Khuat TT, Gabrys B. A comparative study of general fuzzy min-max neural networks for pattern classification problems. Neurocomputing. 2020, 386. https://doi.org/10.1016/j.neucom.2019.12.090
7. Li W, Zhang K, Chen Y, Tang C, Ma X, Luo Y. Random Fuzzy Clustering Granular Hyperplane Classifier. IEEE Access. 2020. https://doi.org/10.1109/ACCESS.2020.3046224
8. Meng X, Liu M, Wu J, Zhou H, Xu F, Wu Q. Hierarchical clustering on metric lattice. International Journal of Intelligent Information and Database Systems. 2020;13(1). https://doi.org/10.1504/IJIIDS.2020.108214
9. Nilashi M, Rupani PF, Rupani MM, Kamyab H, Shao W, Ahmadi H et al. Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach. Journal of Cleaner Production. 2019, 240. https://doi.org/10.1016/j.jclepro.2019.118162
10. Patel J, Yadav RS. Applications of Clustering Algorithms in Academic Performance Evaluation. OALib, 2015;02(08). https://doi.org/10.4236/oalib.1101623
11. Wang D, Yang F, Gan L, Li Y. Fuzzy prediction of power lithium ion battery State of Function based on the fuzzy c-means clustering algorithm. World Electric Vehicle Journal. 2019;10(1). https://doi.org/10.3390/wevj10010001
12. Yadav RS, Ahmed P, Soni AK, Pal S. Academic performance evaluation using soft computing techniques. Current Science. 2014;106(11). https://doi.org/10.18520/cs/v106/i11/1505-1517
13. Yin X. Construction of Student Information Management System Based on Data Mining and Clustering Algorithm. Complexity, 2021. https://doi.org/10.1155/2021/4447045