INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Analyzing Student Performance using Fuzzy Possibilistic C-Means Clustering Algorithm

R Jayasree[1,2]*, N A Sheela Selvakumari[3]

1 Research Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India
2 Assistant Professor, Department of BCA, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India
3 Associate Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India

OPEN ACCESS

Received: 18-07-2023
Accepted: 14-08-2023
Published: 13-10-2023

Citation: Jayasree R, Selvakumari NAS (2023) Analyzing Student Performance using Fuzzy Possibilistic C-Means Clustering Algorithm. Indian Journal of Science and Technology 16(38): 3230-3235. https://doi.org/10.17485/IJST/v16i38.226

* **Corresponding author**.

jsreewin@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This work is to propose a more effective Fuzzy C-means clustering algorithm for predicting student performance based on their health. **Methods:** The standard dataset is collected from UCI repository. This study proposes FPCM-SPP clustering algorithm which is compared with traditional algorithms like K-Means, K-Medoids, and Fuzzy C-Means using student data from secondary education at two Portuguese institutions (2008). Based on the clustering accuracy, mean squared error, and cluster formation time, the performance of the clustering methods is compared. **Findings:** The proposed Fuzzy Possibilistic C-Means for Student Performance Prediction (FPCM-SPP) Algorithm, according to the observational findings, performed the best of all the models. Predicting student's current health status at an early stage of the school can help academia not only to concentrate more on the healthy students but also to apply more efforts in developing remedial programs for the weaker ones in order to improve their progress while attempting to avoid student dropouts. **Novelty:** The innovative aspect of this research is that it suggests ways to improve on the performance of earlier algorithms through modifications to the FPCM's objective function.

Keywords: Prediction; Clustering; K-Means; K-Medoids; Fuzzy C-Means; FPCM-SPP

1 Introduction

The data mining approach that is used most frequently is the clustering. Clustering is a method that is used to organize students into groups that are comparable to one another in terms of their attributes and capabilities[(1)]. Throughout the course of their histories, educational institutions have accumulated huge quantities of data pertaining to teachers, students, and other relevant parties. Unsupervised machine learning techniques are used by researchers, particularly clustering of educational data. Using data from online learning platforms, researchers created clustering algorithms

to forecast student performance[2]. Common elements include student data, e-learning behavior, and learning outcomes. Several Malaysian public university students majoring in a variety of subjects are included in the data collection. Using k-means, BIRCH, and DBSCAN, researchers create models for learning from educational data[3,4]. These techniques have advantages and disadvantages, but they provide good clusters. It is challenging to obtain accurate cluster results because of the algorithm's sensitivity to parameter changes.

The practice of mining data for research purposes in the realm of education is gaining prominence. The implementation of data mining techniques comprises the design of methods that extract information from data and are used to use data to unearth information that has been disguised. The information that was obtained can be put to use to better understand student behavior, aid teachers, improve instruction, evaluate and enhance e-learning systems, enhance student academic achievement, improve curriculum, and give a variety of other benefits. Križanić[5] also investigated how student conduct in an e-learning environment can influence exam performance. Decision trees and k-means are used in data mining. Three decision tree models were built for three student groups with similar e-learning patterns based on the cluster analysis. The biggest improvement in knowledge came from midterm tests. Exam performance would decline with less lecture and online instruction.

1.1 Research Gap

The purpose of this research is to identify the clusters which are based on similar values. In previous works, traditional clustering algorithms are applied to predict the student performance. Since novel algorithm and framework is required for prediction.

2 Methodology

2.1 Dataset Description

In this dataset, 33 attributes are used in each record to characterize student achievement in secondary education of two Portuguese schools. The attributes include student grades, demographic, social, and educational factors. The data was collected through school reports and surveys. Two datasets are presented in relation to the performance in the two distinct subjects of Portuguese language (por) and mathematics (mat). The dataset contains student's current health status (numeric: from 1 — very bad to 5 - very good).

The student's health status is predicting using proposed fuzzy possibilistic clustering algorithm and the proposed algorithm K-Means, K-medoids, and Fuzzy C-Means.

2.2 The Proposed Fuzzy Possibilistic Clustering Algorithm for Student Performance Prediction (FPCM — SPP)

$$J_{FCM}(V,U,X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j}^{m} d^2(x_j, v_i), 1 < m < +\infty \tag{1}$$

The student performance dataset is represented as $X = \{x_1, x_2, \ldots\ldots x_n\} \subseteq R^p$ in the first equation. The number of data items is denoted by p, and the number of clusters is denoted by c with a value of $2 \leq c \leq n-1$. v_i is the p-dimension center of the cluster i, whereas $d^2(x_j, v_i)$ is a measure of the distance between object x_j and the cluster center v_i. $V = \{v_1, v_2, \ldots\ldots v_c\}$ is the c centers or prototypes of the clusters. $U = \{u_{ij}\}$ depicts a fuzzy partition matrix where x_j i is the j^{th} of p-dimensional measured data and $u_{ij} = u_i(x_j)$ is the degree of participation of x_j in the i^{th} cluster. It is obvious that the fuzzy partition matrix:

$$0 \prec \sum_{j=1}^{n} u_{i,j} \prec n, \forall i \varepsilon \{1, \ldots, c\} \tag{2}$$

$$\sum_{i=1}^{c} u_{ij} = 1, \forall i \varepsilon \{1, \ldots, n\} \tag{3}$$

where, m is a weighted exponent parameter that is applied to each fuzzy membership and determines the level of fuzziness of the final classification; it is a constant value that is greater than one. It is possible to get the minimum value of the objective function J_{FCM} when U is a constraint. To be more specific, taking J_{FCM} with regard to u_{ij} and v_j and zeroing them correspondingly are required situations for J_{FCM} to be at its local dissipations, but they are not sufficient circumstances. The results of this analysis will be as follows:

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)}\right)^{2/(m-1)}\right], 1 \leq i \leq c, 1 \leq j \leq n \tag{4}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k, 1 \leq i \leq c}{} \tag{5}$$

Due to the fact that the relationships of FCM may not constantly precisely reflect the step to which the data belong in a noisy environment, this may cause false conclusions to be drawn. The reason for this is due to the fact that the actual data will almost definitely contain some abnormalities. The formula of the fuzzy c-partition was made more permissive in order to achieve a possibilistic type of membership function in order to make up for this weakness in FCM, and PCM for unsupervised clustering was recommended as a cure. This was done in order to achieve a possibilistic type of membership function in order to make up for this defect in FCM. Each cluster in the PCM strategy is independent of the others, which means that the component that is formed by the PCM corresponds to a dense area in the dataset, and each cluster in the PCM strategy can be considered to stand alone. The objective function of the PCM can be expressed through the usage of the following statement.

$$J_{PCM}(V,U,X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ik}^{m} d^2(x_j, v_i) + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1-u_{ij})^m \tag{6}$$

$$\eta_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} \|x_j - v_i\|^2}{} \tag{7}$$

$$u_{ij} = \frac{1}{1 + \left[\frac{d^2(x_j, v_i)}{\eta_i}\right]^{\frac{1}{m-1}}} \tag{8}$$

It is a weighting parameter also known as the probabilistic parameter. The value of a weighting parameter known as the probabilistic typicality of the training sample x_j that is associated with the cluster i is given by the notation $x \in [1, \infty]$. The PCM, like most other cluster approaches, depends on the execution of an initialization function. Since each data point is only identified as a member of one cluster at a time rather than as members of all the clusters simultaneously when PCM techniques are used.

The properties of fuzzy and probabilistic C-means have been integrated into a single set of traits. The correct characteristic of the data, which depends on both memberships and typicalities, can be described as follows, depending on which memberships and typicalities are taken into consideration:

$$J_{FPCM}(U,T,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} \left(u_{ij}^{m} + t_{ij}^{\eta}\right) d^2(x_j, v_i) \tag{9}$$

Subject to the following restrictions:

$$\sum_{i=1}^{n} u_{ij} = 1, \forall j \varepsilon \{1, \ldots, n\} \tag{10}$$

$$\sum_{j=1}^{n} t_{ij} = 1, \forall j \varepsilon \{1, \ldots, c\} \tag{11}$$

The objective function is composed of two expressions, the first of which uses a fuzziness weighting exponent for the fuzzy function and the second of which employs a usual weighting exponent for the possibility function. The objective function's two coefficients, however, are solely utilised as indicators of membership and typicality. A new connection that is just a little bit different permits a more rapid drop in the function and an increase in membership and typicality. They raise this degree when they trend towards 1 and decrease it when they trend towards 0. The weighted exponent will be added to the relationship as a distance exhibitor for the two objective functions being considered.

An answer to the problem of maximizing the value of the objective function can be arrived at through an iterative process. During this stage of the procedure, the degrees of membership and typicality, as well as the cluster centers, are modified by employing the equations that are described in the section that comes after this one.

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)}\right)^{2/(m-1)}\right], 1 \leq i \leq c, 1 \leq j \leq n \tag{12}$$

$$t_{ij} = \left[\sum_{k=1}^{n} \left(\frac{d^2 (x_j, v_i)}{d^2 (x_j, v_k)} \right)^{2/(\eta-1)} \right], 1 \leq i \leq c, 1 \leq j \leq n \tag{13}$$

$$v_i = \frac{\sum_{k=1}^{n} \left(u_{ik}^m + t_{ik}^\eta \right) X_K, 1 \leq i \leq c}{} \tag{14}$$

Along with the conventional point prototypes or cluster centers for each group, FPCM simultaneously creates memberships and possibilities. The FPCM is a hybrid of PCM and FCM that often avoids the numerous problems that are associated with PCM, FCM, and the FPCM. FPCM eliminates the noise sensitivity issue that outbreaks FCM and triumphs over the coincident clusters issue that outbreaks PCM. The estimation of centroids, on the other hand, is affected by the noise data. Because of this challenge, the FPCM-SPP that has been proposed has been constructed.

3 Results and Discussion

The experiment was conducted with some fine-tuning of the parameters. The information was gathered from student reports and questionnaires submitted by the schools. Two different sets of data can be obtained here, one for assessing proficiency in the Portuguese language and the other for measuring success in mathematics. To evaluate the proposed FPCM-SPP and exiting K-Means, K-Medoids, Fuzzy C-Means algorithms with clustering accuracy, mean absolute error, and execution time.

Table 1 Illustrates the comparison of existing algorithms for student performance prediction using clustering algorithms in educational data mining.

Table 1. Comparison of existing algorithms for Educational Dataset

References	Model Descriptions	Results
Hooshyar et al. (2020) [2]	Proposed method to predict learning disability student performance using procrastination behaviour (PPP)	Achieved 96% accuracy compare than L-SVM.
Jiashun et al. (2021) [3]	Proposed a new possibilistic fuzzy C-means based on the weight parameter algorithm (WPFCM)	WPFCM resubstituting errors are slightly lower than FCM and PFCM, but significantly lower than PCM. WPFCM offers a fast speed of convergence and requires minimal running time for huge datasets.
Valarmathy et al. (2019) [6]	Hierarchical methods, Partitioning methods, and Density-based methods	DBSCAN algorithm achieves 0.9675 accuracy.
Vital et al. (2019) [7]	K-means and hierarchical cluster statistical Analysis	The Hierarchal Cluster Study hypothesises the causes of student success and failure based on a variety of variables.
Qiu et al. (2022) [8]	Proposed the process-behaviour classification (PBC) model, an online behaviour classification approach based on the e-learning process.	The F1-score, Kappa value, and prediction performance of the PBC model are all higher.

The Figure 1 shows the proposed FPCM-SPP algorithm achieves 92.4% accuracy for Portuguese lesson dataset and 6.2% greater than Fuzzy C-Means, 8.1% greater than K-Mediods, and 9.7% greater than K-Means algorithm. The proposed FPCM-SPP algorithm obtain 93% accuracy for mathematics lesson dataset and 6.8% greater than Fuzzy C-Means, 10.2% greater than K-Mediods, and 11.5% greater than K-Means algorithm.

Figure 2 shows the proposed FPCM-SPP algorithm achieves 0.3895% Mean Squared Error for Portuguese lesson dataset and 0.0957% less than Fuzzy C-Means, 0.1061% less than K-Mediods, and 0.0725 % less than K-Means algorithm. The proposed FPCM-SPP algorithm obtain 0.3652% accuracy for mathematics lesson dataset and 0.0604% less than Fuzzy C-Means, 0.0473 % less than K-Mediods, and 0.0459% less than K-Means algorithm.

The Figure 3 shows the proposed FPCM-SPP algorithm takes 0.30 seconds time taken execution of Portuguese lesson dataset and the proposed FPCM-SPP algorithm obtain 0.28 seconds time taken execution of mathematics lesson dataset.

The Figure 4 shows that the proposed FPCM-SPP clustering algorithm obtains greater accuracy, less mean squared error, and minimum execution time. Hence, the dataset is clustered based on student's current health status.

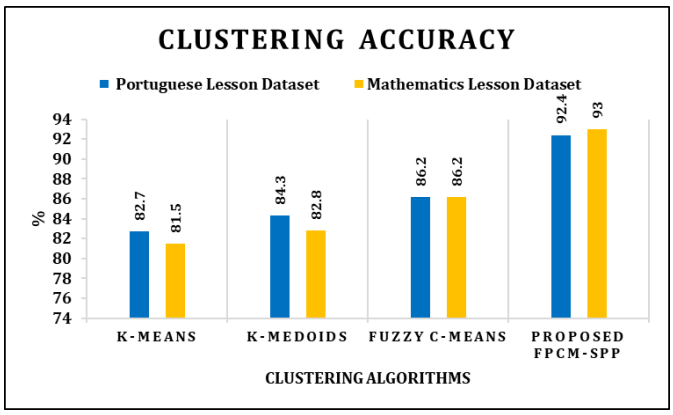

Fig 1. Clustering Accuracy

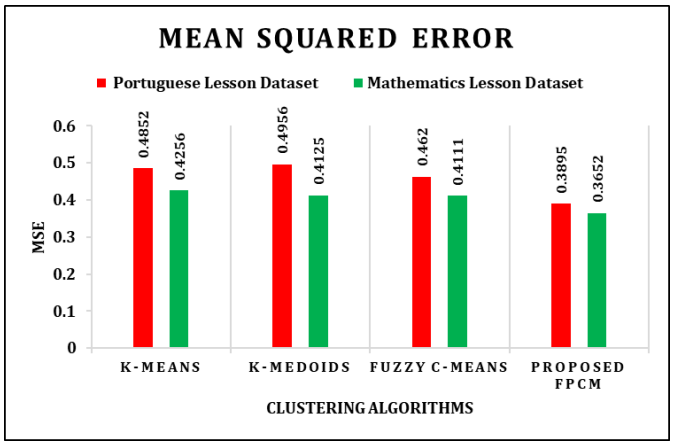

Fig 2. Mean Squared Error

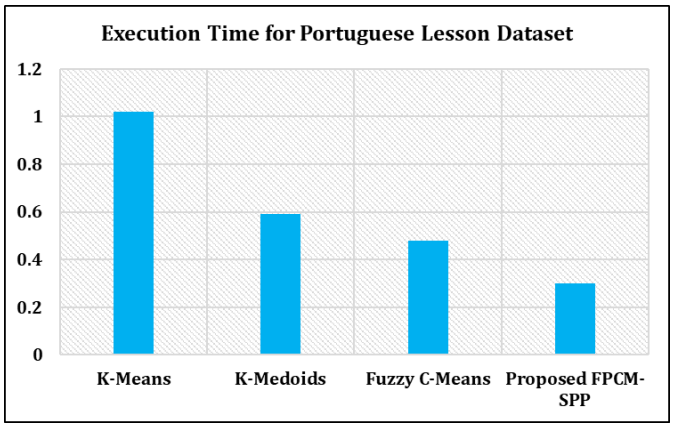

Fig 3. Mean Squared Error for Portuguese Lesson

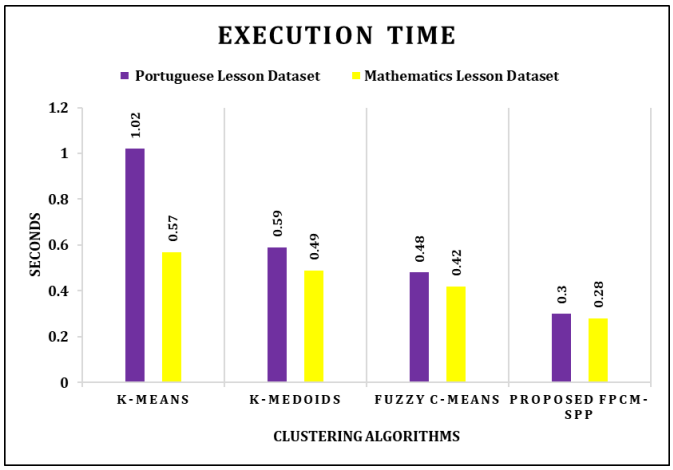

Fig 4. Execution Time

4 Conclusion enrich with quantitative data

This study has proposed FPCM-SPP clustering algorithm for predicting students' performance based on their health. The novelty of this work lies in modified objective function in FPCM clustering algorithm. The proposed FPCM-SPP is the hybridization of FCM and PCM. According to the findings, the FPCM-SPP algorithm that was proposed has superior performance when compared to the other three clustering algorithms. The proposed FPCM-SPP algorithm achieves 92.4% accuracy for por dataset and 93% accuracy for mat dataset. FPCM-SPP performs with 0.3895 and 0.3652 mean squared error for por dataset and mat dataset. Additionally, the execution time of the proposed algorithm is 0.3 seconds for por dataset and 0.28 seconds for mat dataset. The limitation of this research work is comparing the proposed algorithm to traditional existing algorithms like K-Means, K-medoids, Fuzzy C-Means and the dataset had limited features. In the future, the dataset should be expanded by including additional characteristics associated to student performance, behaviour, mental health, and employment, and the clustering performance should be further improved by applying the right feature selection approaches and various parameter modifications. Furthermore, the proposed algorithm is to compare with the optimization algorithms for analysing its performance.

References

1) Gu Y, Ni T, Jiang Y. Deep Possibilistic C-means Clustering Algorithm on Medical Datasets. *Comput Math Methods Med*. 2022;2022:1–10. Available from: https://doi.org/10.1155/2022/3469979.
2) Hooshyar D, Pedaste M, Yang Y. Mining Educational Data to Predict Students' Performance through Procrastination Behavior. *Entropy*. 2020;22(1):1–24. Available from: https://doi.org/10.3390/e22010012.
3) Chen J, Zhang H, Pi D, Kantardzic M, Yin Q, Liu X. A Weight Possibilistic Fuzzy C-Means Clustering Algorithm. *Scientific Programming*. 2021;2021:1–10. Available from: https://doi.org/10.1155/2021/9965813.
4) Križanić S. Educational data mining using cluster analysis and decision tree technique: A case study. *International Journal of Engineering Business Management*. 2020;12(January-December):1–9. Available from: https://doi.org/10.1177/1847979020908675.
5) Li X, Zhang Y, Cheng H, Zhou F, Yin B. An Unsupervised Ensemble Clustering Approach for the Analysis of Student Behavioral Patterns. *IEEE Access*. 2021;9:7076–7091. Available from: https://doi.org/10.1109/ACCESS.2021.3049157.
6) Valarmathy N, Krishnaveni S. Performance Evaluation and Comparison of Clustering Algorithms Used in Educational Data Mining. *International Journal of Recent Technology and Engineering (IJRTE)*. 2019;7(6S5):103–112. Available from: https://www.ijrte.org/wp-content/uploads/papers/v7i6s5/F10180476S519.pdf.
7) Vital TPR, Lakshmi BG, Rekha HS, Dhanalakshmi M. Student Performance Analysis with Using Statistical and Cluster Studies. In: Nayak J, Abraham A, Krishna B, Sekhar GC, Das A, editors. Soft Computing in Data Analytics;vol. 758 of Advances in Intelligent Systems and Computing. Springer. 2018;p. 743–757. Available from: https://doi.org/10.1007/978-981-13-0514-6_71.
8) Qiu F, Zhang G, Sheng X, Jiang L, Zhu L, Xiang Q, et al. Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*. 2022;12(453):1–15. Available from: https://doi.org/10.1038/s41598-021-03867-8.