Comparison between single and multi-objective clustering algorithms: MathE case study *

Beatriz Flamia Azevedo[1,2,4][0000−0002−8527−7409], Ana Maria A. C. Rocha[2][0000−0001−8679−2886], Florbela P. Fernandes[1,4][0000−0001−9542−4460], Maria F. Pacheco[1,3,4][0000−0001−7915−0391], and Ana I. Pereira[1,2,4][0000−0003−3803−2043]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Bragança - 5300-253, Portugal
[2] ALGORITMI Research Centre / LASI, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[3] Center for Research & Development in Mathematics and Applications CIDMA, University of Aveiro, Aveiro, Portugal
[4] Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Bragança - 5300-253, Portugal
{beatrizflamia, fflor, pacheco, apereira}@ipb.pt, arocha@dps.uminho.pt

Abstract. This paper compares the results obtained for four single clustering algorithms with a multi-objective clustering approach. For this, a dataset describing the student's behavior within the Linear Algebra topic on the MathE e-learning platform is used. This dataset aids in understanding student performance and engagement in MathE to support the development of an intelligent system to tailor the platform's resources to users's needs. The four algorithms suggested two clusters as the optimal solution for the dataset. However, this binary categorization did not provide meaningful insights into the proposal of the MathE platform; that is, it did not provide a customized system according to individual needs. Thus, this study uses the multi-objective clustering algorithm, which results in a set of non-dominated solutions, providing decision-makers with a broader range of options to choose the solution that best meets their needs. The results demonstrate the main benefits of the proposed human-in-the-loop multi-objective approach since it provides several optimal solutions and allows the decision-maker to apply fundamental knowledge to define the most appropriate solution to the problem based on previous knowledge.

Keywords: multi-objective clustering · automatic clustering · optimization · bio-inspired algorithm.

* This work has been supported by FCT Fundação para a Ciência e Tecnologia within the R&D Units Project Scope UIDB/00319/2020, UIDB/05757/2020 (DOI: 10.54499/UIDB/057 57/2020), UIDP/05757/2020 (DOI: 10.54499/UIDP/05757/2020) and Erasmus Plus KA2 within the project 2021-1-PT01-KA220-HED-000023288. Beatriz Flamia Azevedo is supported by FCT Grant Reference SFRH/BD/07427/2021.

1 Introduction

Clustering is a data analysis technique used to partition a set of objects into groups (clusters) based on measures of similarity [17]. The goal of clustering is to ensure that elements within the same cluster are more similar to each other than they are to elements in different clusters. This technique is commonly used in several fields, such as machine learning, data mining, pattern recognition, and image analysis, to discover hidden patterns or structures within data [5,17].

With several methods and algorithms available in the literature, choosing the most suitable one for each problem becomes challenging [5]. Ideally, testing all these methods would be the most effective approach, but this would require considerable time, which is often unavailable. Partitioning clustering algorithms are guided by internal and external measures. The selection of these measures plays a crucial role in the algorithm's response. Given the wide range of measures available [17,18,19], choosing the ideal one for each dataset represents a significant challenge. Thus, employing collaborative strategies represents a substantial advantage in addressing it. Combining several measures within the same approach and adopting a multi-objective approach expands the range of solution options, providing greater variety and flexibility in decision-making [4].

Multi-objective optimization is an area of multiple-criteria decision-making concerning mathematical optimization problems that involve several objective functions to be minimized or maximized simultaneously. These objective functions are conflicting, meaning there is a trade-off between them [9]. An effective objective function is crucial for the success of single-objective clustering. However, different datasets may need distinct metrics (objective functions) for optimal performance. Given the abundance of conflicting metrics to evaluate clustering effectiveness, it is necessary to simultaneously optimize multiple measures to capture the varied characteristics inherent in different datasets [11]. Thus, as Ferligoj and Batagelj [12] pointed out, clustering represents a multi-objective optimization problem.

While the single-objective approach aims to find a single optimal solution for a single criterion, the multi-objective approach seeks to optimize multiple criteria simultaneously, resulting in a set of non-dominated solutions (Pareto front solutions) [5]. This gives decision-makers a broader range of options to choose the solution that best meets their needs. In complex problems with trade-offs between different objectives, the multi-objective approach can lead to more informed decisions and more robust results.

In cluster analysis, there may be cases where a cluster measure defines that an element is similar to another and are in the same cluster, but for other measures, the same elements are not similar and, therefore, are in different clusters. Additionally, there may be situations where the decision-maker has crucial information about the data, but this information is difficult to model mathematically and incorporate into the problem modeling. Thereby, integrating the human-in-the-loop strategy, allowing the decision-maker to choose from a set of solutions (Pareto front solutions) means providing the opportunity to include additional and fundamental knowledge for the problem's response [4].

In the work [3], four clustering single-objective algorithms were used to evaluate the student's behavior when answering questions under the Linear Algebra topic in the MathE platform. MathE (mathe.pixel-online.org) is an e-learning platform, launched in 2019, where students can study mathematics autonomously. In the platform, any student worldwide has free access to resources such as videos, training tests, and pedagogical materials covering several areas of mathematics taught in higher education courses. Within the scope of iMath project (imath.pixel-online.org), the MathE Platform is being improved to become even more interactive and customized according to the user's individual needs. The collected data since the MathE platform is online is being analyzed to recognize patterns and enhance the students' experience while users.

The objective of the work [3] was to group a set of MathE students according to their performance, whether they are getting the correct or incorrect answers when answering the questions available at the Linear Algebra topic, and also consider their engagement on the platform through the number of questions answered. To evaluate the dataset, the work [3] used three automatic single-objective clustering based on different bio-inspired approaches (Particle Swarm Optimization, Genetic Algorithm, Differential Evolution) [20] and with the k-means algorithm [15], which is a non-automatic single-objective clustering algorithm. All four algorithms suggested that two clusters were the optimal solution. However, categorizing the students into two groups does not give the project managers significant conclusions about the student's behavior. Thereby, this study aims to compare the results obtained in [3] with a new algorithm called Multi-objective Clustering Algorithm (MCA), proposed in [4] and described in section 3.1.

Unlike the four single-clustering algorithms used in [3], the MCA is a multi-objective clustering algorithm that integrates several clustering measures to identify the optimal set of solutions (Pareto front solutions). Furthermore, the MCA uses a collaborative strategy involving several cluster validity indexes to tackle the issue of selecting a solution from the Pareto front when the decision-maker judges necessary.

This paper is organized as follows: after the introduction, Sect. 2 introduces the clustering concepts and some methods and strategies of data categorization. Afterwards, Sect. 3 presents some clustering measures and defines the Multi-objective Clustering Algorithm used in this paper. Section 5 presents and discusses the obtained results. Finally, the main conclusion and future direction are given in Sect. 6.

2 Clustering Algorithms

Clustering algorithms can be categorized into families based on their characteristics. The optimization strategy employed can be either single-objective or multi-objective. Regardless of the optimization strategy, the clustering algorithms divide into crisp or fuzzy algorithms, further partitioning into k-based or automatic clustering methods, as presented in Figure 1.

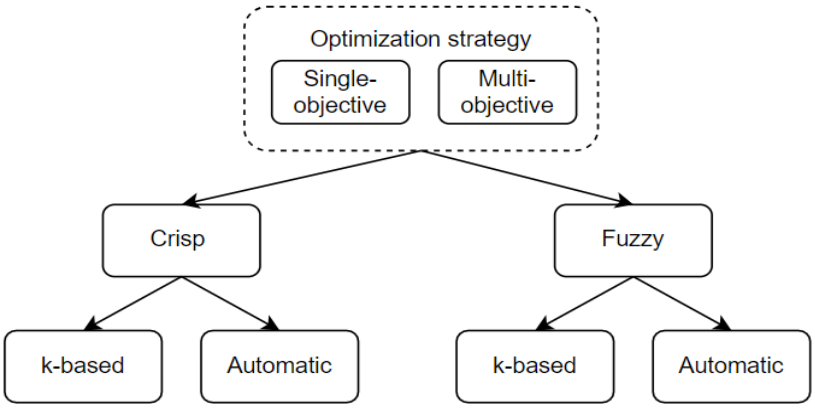


Fig. 1: Clustering algorithms strategies

In the case of a single-objective optimization problem, there is only one objective function to be optimized, and its superiority is determined by comparing the objective function values. Meanwhile, in the multi-objective optimization problem, the problem to be optimized involves multiple conflicting objectives, i.e., a vector of objective functions. In this case, the quality of a solution is determined by the dominance criterion [9].

Crisp clustering assigns each data point to exactly one cluster, with clear boundaries between cluster frontiers. Thus, each element is associated with a cluster centroid that is closest to in terms of a given clustering measure. On the other hand, fuzzy clustering is very useful in the presence of uncertainty or ambiguity in data features [1,5]. This method allows a data element to belong to multiple clusters simultaneously with varying degrees of membership [1]. It assigns membership values to data elements, indicating the degree to which they belong to each cluster. Unlike crisp clustering, in fuzzy clustering, there are no strict boundaries between clusters; instead, there is a gradual transition in membership values, so each data element contributes to each cluster based on its similarity to the cluster centroid [1].

Automatic clustering algorithms can group data into clusters without requiring prior knowledge of the number of clusters. These algorithms are designed to automatically determine the optimal cluster configuration based on the inherent structure of the data. These algorithms typically employ statistical measures, optimization techniques, or heuristics to determine the optimal number of clusters and assign data points to clusters [1]. In turn, k-based algorithms require the k from the user, which is a difficult parameter to predict, that can be chosen by the decision based on previous knowledge about the data or by selecting a validity measures criterion [15]. Consequently, the selection of this measure determines the quality of clusters [11].

3 Multi-objective Clustering Method

The following sections present the bi-objective problem, the clustering measures used, and the Multi-objective Clustering Algorithm. In the end, the cluster collaborative indices procedure is presented.

3.1 Multi-objective Problem

This study explores a bi-objective programming problem whose objective functions are based on intra-clustering and inter-clustering measures. The bi-objective problem can be defined as the simultaneous minimization of an intra-clustering measure and the maximization of an inter-clustering one, as formulated in (1):

$$\min \{f_a, -g_b\} \tag{1}$$

where f_a, for $a = 1, 2$, is an intra-clustering measure and g_b, for $b = 1, ..., 6$, is the inter-clustering measure, in which a and b represent a measure, among the ones described bellow.

3.2 Clustering Measures

To cluster the dataset into distinct sets, it is crucial to establish criteria for computing distances between individual elements [18,19]. This research investigates conventional measures and potential modifications in intra- and inter-clustering metrics, detailed below. Throughout this study, the Euclidean distance is denoted by $D(\cdot, \cdot)$.

Intra-clustering Measures refer to the distance between elements within a specific cluster. Several methods exist for computing intra-cluster measures, and the ones explored in this study are outlined below.

The measure $SAxc$ refers to the sum of average distance between the elements to their centroid in terms of the number of elements belonging to each cluster set $\#C_j$, where Sxc_j represents the sum of the distances between the elements and the centroid j as defined in Equation 2.

$$SAxc = \sum_{j=1}^{k} \frac{Sxc_j}{\#C_j} \tag{2}$$

Another inter-clustering measure that can be used is the FNc, the sum of the distances of the furthest neighbor within the cluster. It evaluates the sum of the furthest neighbor distance of each cluster C_j, where x_i^j, and x_l^j belong to the same cluster j.

$$FNc = \sum_{j=1}^{k} \max\{D(x_i^j, x_l^j)\} \quad \text{for} \quad i = 1, ..., \#C_j, \; l = 1, ..., \#C_j, \; i \neq l. \tag{3}$$

Inter-clustering measures quantify the distance between elements belonging to different clusters. The inter-cluster measures considered are detailed below.

The measure Scc represents the sum of the distances between centroids [18], and is defined in Equation 4,

$$Scc = \sum_{\substack{t,j=1,\\ t\neq j}}^{k} D(c_t, c_j). \tag{4}$$

Thus, the average of the distances between centroids (Acc) is based on Scc, and it is presented in Equation 5,

$$Acc = \frac{Scc}{k}. \tag{5}$$

Another measure considered in the $FNcc$, which is the sum of the furthest neighbor distance between elements of different clusters C_j, known as complete linkage [19] as shown in Equation 6,

$$FNcc = \sum_{j=1}^{k}\sum_{t>j}^{k} \max\{D(x_i^j, x_l^t)\} \text{ for } i = 1, ..., \#C_j, l = 1, .., \#C_t, i \neq j. \tag{6}$$

Thus, the average of the sum of the furthest neighbor distances between elements of k different clusters ($AFNcc$) is described in Equation 7,

$$AFNcc = \frac{FNcc}{k}. \tag{7}$$

Finally the sum of the nearest neighbor distance between elements of different clusters ($NNcc$), also known as single linkage [18], is defined in Equation 8,

$$NNcc = \sum_{j=1}^{k}\sum_{t>j}^{k} \min\{D(x_i^j, x_l^t)\} \text{ for } i = 1, ..., \#C_j,\ l = 1, .., \#C_t, i \neq l. \tag{8}$$

And, the average of the sum of the nearest neighbor distance between elements of k different clusters ($ANNcc$) is defined in Equation 9,

$$ANNcc = \frac{NNcc}{k}. \tag{9}$$

3.3 Multi-objective Clustering Algorithm

The Multi-objective Clustering Algorithm (MCA) evaluates both intra-cluster and inter-cluster measures to autonomously define the optimum number of clusters [4,6]. This is achieved by minimizing distances within clusters and maximizing distances between clusters, integrating different combinations of intra- and inter-clustering measures throughout the algorithm execution.

To explain the MCA, consider a dataset $X = \{x_1, x_2, ..., x_m\}$ consisting of m elements, where $x_i \in \mathbb{R}^d$ (with d being the number of variables in the dataset). The goal is to divide X into k optimal groups (clusters). Since MCA can autonomously determine the optimal number of clusters, it needs to define the range of potential partitions. This entails specifying the minimum k_{min} and maximum k_{max} number of centroid possibilities.

The algorithm starts by receiving the dataset as input. Next, a pair of measures, consisting of one intra-measure and one inter-measure, is automatically selected from those outlined in Subsection 3.2. Using this choice, the MCA employs the Multi-objective Particle Swarm Optimization (MOPSO) [8], to generate the Pareto front associated with the bi-objective function of the problem. This process occurs until all combinations of measure pairs have been evaluated. Upon completing this procedure, a total of 12 distinct Pareto fronts is obtained.

A specific pair of measures generates each Pareto front. Due to the diverse magnitudes of the measures considered, it is necessary to normalize the solutions that compose each Pareto front to ensure a fair comparison of solutions. Consequently, every Pareto front undergoes normalization through the Min-Max scaling method [16]. In this process, each solution s on the Pareto front is individually normalized to the range $[0, 1]$, where s_{min} and s_{max} represent the smallest and largest solution values within the respective Pareto front. Subsequently, all normalized solutions across the Pareto fronts are evaluated based on dominance [9], with non-dominated solutions being selected to construct the Hybrid Pareto front (HPF). This HPF is the output of the MCA algorithm, comprising the set of non-dominated solutions derived from all solutions obtained from the Pareto fronts associated with each pair of measures.

3.4 Cluster Collaborative Indices Procedure

Cluster Validity Indices (CVI) assess clustering quality by considering intra-cluster cohesion and inter-cluster separation. They aim to distinguish between superior and inferior solutions, ensuring clustering algorithm efficiency. An optimal solution for one specific CVI could not be optimal for another CVI [14]. To evaluate the quality of the solutions generated by the Hybrid Pareto front, the Cluster Collaborative Indices Procedure (CCIP) was developed. In this procedure, each solution of the Hybrid Pareto front was evaluated by each CVI criterion. Therefore, the solution that receives more indications is considered the best to be selected by the decision-maker. In case of a tie, the set of solutions indicated are considered as the most appropriate for the problem, and it is up to the decision-maker to make the final decision. The CVI used in this work are described below.

3.5 Davies-Bouldin Index

The Davies-Bouldin index (DB) [2], estimates the cohesion based on the distance from the elements x_i^j in a cluster to its centroid c_j and the separation based on the distance between centroids $D(c_j, c_t)$. The smallest DB indicates the optimal

partition. First, it is necessary to evaluate an intra-cluster measure represented by the mean distance between each element within the cluster x_i^j and its centroid c_j, which is a dispersion parameter $S(c_k)$, as Equation 10,

$$S(c_j) = \sum_{i=1}^{\#C_j} \frac{D(x_i^j, c_j)}{\#C_j}. \tag{10}$$

And, the DB index is given by Equation 11,

$$DB = \frac{1}{k} \sum_{j=1}^{k} \max_{\substack{t=1,\dots,k, \\ j \neq t}} \left\{ \frac{S(c_j) + S(c_t)}{D(c_j, c_t)} \right\}. \tag{11}$$

3.6 Dunn Index

The Dunn index (DI) [10] is a ratio-type index where the cohesion is estimated by the nearest neighbor distance and the separation by the maximum cluster diameter. Thus, a higher DI will indicate compact, well-separated clusters, while a lower index will indicate less compact or less well-separated clusters [10]. The higher the value of the Dunn index, the better the quality of the clusters. So, DI is defined as the rate between the minimum distance between elements of different clusters, and the largest distance between elements of the same cluster (sometimes called cluster diameter), as in Equation 12.

$$DI = \frac{\min\limits_{\substack{j,t=1,\dots,k, \\ j \neq t}} \{D(x_i^j, x_l^t)\}}{\max\limits_{j=1,\dots,k} \{D(x_i^j, x_l^j)\}} \tag{12}$$

3.7 Calinski-Harabasz Index

The Calinski-Harabasz (CH) [7] is a ratio-type index where the cohesion is estimated based on the distance from the elements in a cluster to its centroid [2,7]. First, it is necessary to calculate the inter-cluster dispersion (BGSS), which measures the weighted sum of squared distance between the centroids of a cluster, c_j, and the centroid of the whole dataset, denoted as $\overline{X}$ which represents the barycenter of the X dataset, as represented in Equation 13,

$$\overline{X} = \frac{1}{m} \sum_{i=1}^{m} x_i. \tag{13}$$

The BGSS is defined as Equation 14

$$BGSS = \sum_{j=1}^{k} \#C_j \times D^2(c_j, \overline{X}). \tag{14}$$

The second step is to calculate the intra-cluster dispersion for each cluster j, evaluated by the sum of the squared distance between the elements x_i^j until its centroid c_j, as defined in Equation 15.

$$WCSS = \sum_{j=1}^{k} \sum_{i=1}^{\#C_j} D^2(x_i^j, c_j) \tag{15}$$

The CH index seeks to maximize the dispersion between clusters (increasing BGSS) while minimizing the dispersion within clusters (decreasing WCSS). The higher the value of the CH index, the better the separation between clusters. Thus, the CH index is defined as Equation 16:

$$CH = \frac{m-k}{k-1} \times \frac{BGSS}{WCSS} \tag{16}$$

4 Dataset

The MathE platform aims to offer a dynamic and motivating way of teaching and learning mathematics, relying on interactive digital technologies that enable autonomous study [3]. The primary objective of this study is to compare the performance of the four single-optimization algorithms used in [3] with the performance of the MCA. In addition, it intends to explore the characteristics of students through the application of a clustering algorithm, aiming to identify patterns among platform users. These patterns will be leveraged to enhance the platform's intelligence, enabling personalized responses to questions based on individual student profiles.

For this, the same dataset used in [3] was considered. This dataset is composed of the information of the students who performed Linear Algebra topics in the MathE platform. This topic is considered since it is the most accessed topic on the platform. The dataset comprises the information of 3696 answers across 199 distinct questions that 134 students across several countries answered. Thereby, the dataset consists of 142 basic and 69 advanced questions. Table 1 details the correct and incorrect answers based on the question level. Of the total responses, 2919 were at the basic level and 777 at the advanced level, resulting in 1741 and 1955 correct and incorrect answers, respectively.

Each student's correct and incorrect answers were evaluated across basic and advanced levels for clustering in order to analyze students' profiles. Outlier students were identified using the Box plot method and subsequently removed from the dataset, resulting in data from 99 students available for analysis.

Similar to the work [3], to perform the clustering analysis, 4 variables were defined, as presented below. Each of them describes the number of questions a student answers according to the question difficult level and the type of answer.

- variable 1: correct answers to basic questions
- variable 2: incorrect answers to basic questions
- variable 3: correct answers to advanced questions
- variable 4: incorrect answers to advanced questions

Table 1: Number of questions answered according to the type of answer given

	Answer Type		
Level	Correct	Incorrect	Total
Basic	1386	1533	2919
Advanced	355	422	777
Total	1741	1955	3696

5 Results and Discussion

In the work [3], the dataset presented was analyzed using three automatic single-objective clustering algorithms based on different metaheuristics (Particle Swarm Optimization, Genetic Algorithm, Differential Evolution) and also by the k-means algorithm. The findings suggested that partitioning the dataset into two clusters was optimal in both algorithms. However, this solution did not adequately represent the problem and failed to offer meaningful insights into the problem.

Although the four algorithms were able to categorize the students into two different clusters, the proposed solutions allow for few conclusions to be drawn about student behavior, resulting in an ineffective solution for categorizing student behavior. Since the MathE project aims to personalize the platform system according to each user's needs, the clusters must better express each group's similarities and dissimilarities.

While the four algorithms categorized students into two distinct clusters, they could not identify a group with significantly superior performance. A more appropriate approach to address this issue would involve redistributing the data into more clusters using another algorithm until a satisfactory solution is found.

In situations where data analysis reaches high levels of complexity, as is the case under consideration, the results heavily depend on the choice of the algorithm. However, a wide range of algorithms are available in the literature, and choosing the ideal one for each dataset is an extremely challenging task. When performing the data analysis with four different algorithms, although the solutions were optimal from the optimization point of view, each algorithm presented a different solution regarding centroid positions and data distribution. However, they all proposed a division based on two clusters, which is not interesting for the decision-maker. In the case of k-means, a non-automatic algorithm, the decision-maker could use another method to determine the number of clusters and iterate until a solution meets their requirements. However, nothing can be done for the other three automatic algorithms.

In situations like this, the MCA stands out because it is a multi-objective algorithm that allows several combinations of solutions to find optimal solutions for the problem (Hybrid Pareto front). The MCA, with its human-in-the-loop approach, provides the decision-maker with greater variability and flexibility to define the best solution based on their expertise. In this way, the MCA allows

for the incorporation of information that is difficult to model mathematically in selecting the final solution to the problem.

Figure 2 presents the Hybrid Pareto front, composed of 34 solutions, generated by combining the 10 different pairs of measures, as detailed in Table 2.

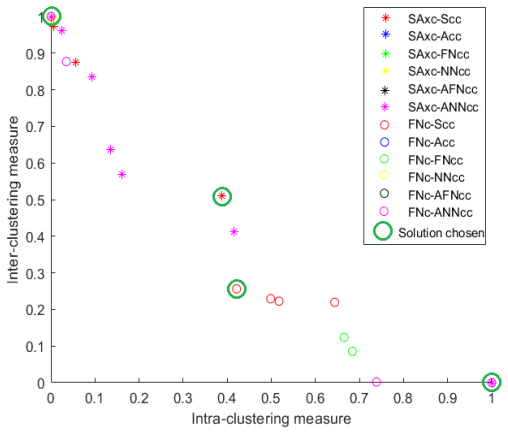


Fig. 2: Hybrid Pareto front

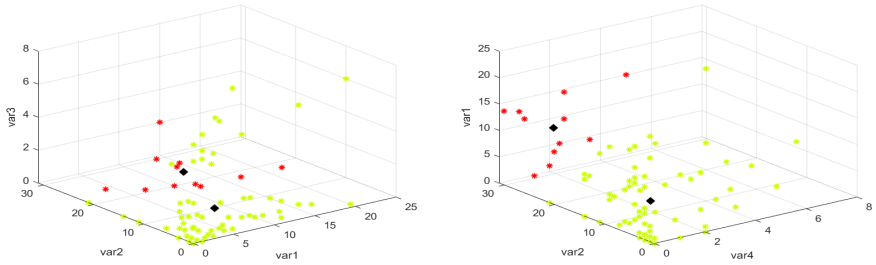
As can be seen in Table 2, MCA provided four possibilities for dividing the dataset into 2, 3, 4, or 5 clusters. Being a multi-objective algorithm, the MCA provides a set of optimal solutions, unlike single-objective algorithms that propose only one. Figures 3 - 6 present four possible divisions suggested by the MCA, considering 2, 3, 4, and 5 clusters, respectively. In all figures, the clusters are presented in a different color, whereas the centroids are black dots.

Table 2: Hybrid Pareto front description

Pareto front Measures	Num. Solutions	Num. Clusters	Pareto front Measures	Num. Solutions	Num. Clusters
SAxc-Scc	4	2 and 3	FNc-Scc	5	2 and 4
SAxc-Acc	0	-	FNc-Acc	1	2
SAxc-FNcc	4	2 and 5	FNc-FNcc	0	-
SAxc-AFNcc	1	2	FNc-AFNcc	2	2
SAxc-NNcc	2	2 and 5	FNc-NNcc	2	2
SAxc-ANNcc	9	2, 3, and 5	FNc-ANNcc	4	2 and 5

The solution presented in Figure 3 is identical to that presented in [3], and they was generated by the pair of measures $FNc - NNcc$. Here the students are categorized into 2 clusters, where the yellow cluster encompasses most of

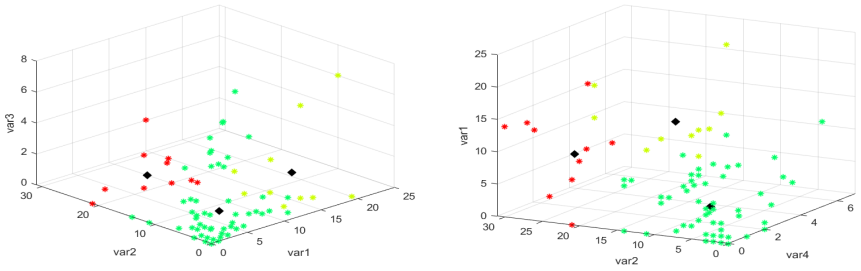
the students, and the red cluster includes a few of them. From this figure, it is possible to see a large concentration of students near the graph's origin in the yellow cluster. This indicates that most students answer to few questions in both difficulty levels. On the other hand, the students who answer more questions are in the red cluster. The results showed that most students choose to answer only basic-level questions, and even within that subset, they make a lot of mistakes. When students risk answering advanced questions, they make even more mistakes, which brings them back to the basic level questions. By analyzing variables 2 and 4 (incorrect answers to basic and advanced questions, respectively), it becomes evident that students in the red cluster answered many questions incorrectly. However, not much can be concluded about the students' behavior as they are categorized in this solution, which only considers two groups of students.



(a) var 1 *vs* var 2 *vs* var 3 (b) var 4 *vs* var 2 *vs* var 1

Fig. 3: MCA solution of two clusters ($k = 2$)

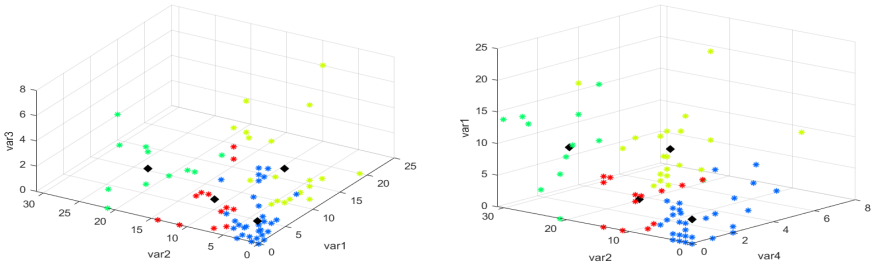
By observing the solution with 3 clusters, presented in Figure 4 and generated by the pair of measures $SAxc - Scc$, it is possible to reach more precise conclusions about the student's profiles. In Figure 4, in the green cluster, there are students who answer a few questions, as they are close to the origin, mainly at the basic level. However, a few students also answer advanced-level questions, with most of their answers being incorrect. In the yellow cluster are the students who answer more basic questions correctly, and some also answer several advanced questions correctly. Finally, in the red cluster are the students who make more mistakes in basic-level questions and answer a few advanced-level questions, most of the time incorrectly. With this solution, it is also possible to compare the clusters. The students in the yellow cluster perform better than those in other clusters, while those in the red cluster perform the worst. Therefore, the green cluster represents an intermediate performance. Our perspective of the dataset changes when moving from an analysis of 2 to 3 clusters. With 3 clusters (Figure 4), it is possible to extract far more information than in the solution with only 2 clusters (Figure 3).



(a) var 1 *vs* var 2 *vs* var 3 (b) var 4 *vs* var 2 *vs* var 1

Fig. 4: MCA solution of three clusters ($k = 3$)

Figure 5 presents a solution with 4 clusters, generated by the pair for measures $FNc - Scc$, enabling interesting conclusions. In the blue cluster, there are the students who answer the fewest questions overall. Regarding the basic level, these students typically answer fewer than 10 questions, and they get a few correct answers. For the advanced-level questions, they answered less frequently, and among the students who attempted to answer them, the majority responded incorrectly. In the red cluster, some students answer a significant portion of basic questions incorrectly, around 10. They answer around 5 basic questions correctly. At the advanced level, they hardly answered at all, and among those who attempted, their performance was very poor. In the yellow cluster, the students correctly answer the highest number of basic questions, with around 20 correct answers and around 5 incorrect ones. These students also answer the advanced level questions but in a smaller quantity than their basic answers. However, at the advanced level, the students in this cluster have the highest accuracy rate compared to those in other clusters (around 3 correct and 6 incorrect). Therefore, in this cluster, there are highly engaged students and those who, most of the time, answer questions correctly. In the green cluster, there are students who,

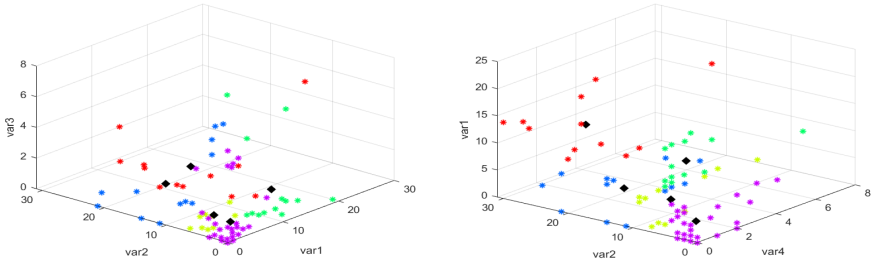


(a) var 1 *vs* var 2 *vs* var 3 (b) var 4 *vs* var 2 *vs* var 1

Fig. 5: MCA solution of four clusters ($k = 4$)

although they answer many basic-level questions, their answers are mostly incorrect. There are around 10 correct answers and 20 incorrect at the basic level. As for the advanced level, they answer a few questions, but most of the answers are incorrect.

Finally, Figure 6 with 5 clusters generated by the measures $FNc - AFNcc$ suggests groupings with few students, allowing for a more detailed analysis. However, considering the dataset comprising only 99 students, considering a decision-maker view, 5 clusters exceed what is expected for the problem, negatively impacting the analysis since the groups of students are too small. However, in situations where more information can be used to define the characteristics of each cluster, a solution that considers 5 clusters may be relevant to the problem.



(a) var 1 *vs* var 2 *vs* var 3 (b) var 4 *vs* var 2 *vs* var 1

Fig. 6: MCA solution of five clusters ($k = 5$)

Table 3 shows the results of the 3 CVI for the 4 solutions previously presented. The Dunn index indicated the solution of 2 clusters, the BD index indicated the solution with 4 clusters as optimal, and the CH index indicated the solution with 3 clusters. Thus, for each CVI, a different solution corresponds to the optimum. Therefore, the decision-maker must select the one that best fits the problem. In the particular case of this dataset, both the solutions with 3 and 4 clusters provide valuable insights into the problem from our point of view as the project decision-maker.

Table 3: Clustering Validity Indices per solution

Clusters per solution	DB index	Dunm index	CH index
2	1.12	**1.73**	29.90
3	0.99	0.05	**69.41**
4	**0.97**	1.29	67.60
5	1.37	1.60	54.70

6 Conclusion

This paper presents a comparison between single and multi-objective clustering algorithms. The results presented in [3], which uses four single-objective clustering algorithms based on the Particle Swarm Optimization, Genetic Algorithm, Differential Evolution [20], and also the k-means algorithm [15], are compared with the Multi-objective Clustering Algorithm [4]. For this, the dataset collected in the MathE platform, considering the students who performed the topic Linear Algebra, was used to investigate the students' grouping patterns, evaluating the performance and engagement to answer the questions.

Multi-objective algorithms consistently deliver solutions equal to or better than single-objective algorithms [13]. The solutions produced by the algorithms used in [3] indicated two clusters as the optimal solution. However, a division into two clusters may not adequately address the problem. In contrast, MCA combines multiple clustering measures to establish the set of optimal solutions (Hybrid Pareto front). This approach offers greater variability in the final solution and increased flexibility for decision-making. The solutions provided by the Hybrid Pareto front allow the selection of the solution that best meets the problem's requirements. In this scenario, as a decision-maker, both solutions with 3 or 4 clusters are suitable for the problem. Additionally, in case of uncertainty, the CCIP strategy could be employed to make the final decision and recommend the solution. As we are dealing with a small problem and only considering four solutions, there is a tie between the CVI results, so each selects one different solution as the best one.

The variability in the generated Hybrid Pareto front highlights the impact of integrating diverse metrics to address a problem. By incorporating multiple measures into the model, exploring a wider array of potential solutions becomes possible. If the model only considers a single pair of measures, it could be restricted to an optimal solution based on that pair, potentially overlooking more suitable alternatives for decision-makers. Therefore, the strategy of employing Hybrid Pareto fronts enriches the final solution of the model. Moreover, unlike methods such as k-means, MCA does not need the prior specification of the number of clusters, addressing a common criticism in the literature [5]. Furthermore, the MCA allows the integration of human expertise, combining the strengths of human intelligence and machine learning algorithms. This integration enables human experts to provide guidance, validate results, and incorporate domain knowledge into the hybrid models [5]. Regarding future developments, it is intends to investigate split and strategies to improve the quality of MCA solutions, mainly in datasets with complex geometries. Regarding the MathE project, exploring other topics covered on the platform must involve other students' features, such as country and course information.

References

1. Aggarwal, C.C., Reddy, C.K.: Data Custering Algorithms and Applications. Taylor & Francis Group. CRC Press (2013)

2. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognition **46**(1), 243–256 (2013)
3. Azevedo, B.F., Rocha, A.M.A.C., Fernandes, F.P., Pacheco, M.F., Pereira, A.I.: Evaluating student behaviour on the mathe platform - clustering algorithms approaches. In: (In press) Book of 16th Learning and Intelligent Optimization Conference - LION 2022. pp. 319–333. Milos - Greece (2022)
4. Azevedo, B.F., Leite, G., Pacheco, M.F., Fernandes, F.P., Rocha, A.M.A.C., Pereira, A.I.: Multi-objective clustering algorithm applied to the mathe categorization problem. Submitted to Information Systems Frontiers
5. Azevedo, B.F., Rocha, A.M.A.C., Pereira, A.I.: Hybrid approaches to optimization and machine learning methods: a systematic literature review. Journal of Machine Learning (2024)
6. Azevedo, B.F., Rocha, A.M.A.C., Pereira, A.I.: A multi-objective clustering algorithm integrating intra-clustering and inter-clustering measures (2024)
7. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974)
8. Coello-Coello, C.A., Lechuga, M.S.: Mopso: a proposal for multiple objective particle swarm optimization. In: Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600). vol. 2, pp. 1051–1056 vol.2 (2002)
9. Deb, K.: Multi-objective Optimization using Evolutionary Algorithms. Wiley & Sons (2001)
10. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics **3**(3), 32–57 (1973)
11. Dutta, D., Sil, J., Dutta, P.: Automatic clustering by multi-objective genetic algorithm with numeric and categorical features. Expert Systems With Applications **137**, 357–379 (2019)
12. Ferligoj, A., Batagelj, V.: Direct multicriteria clustering algorithms. Journal of Classification **9**(1), 43–61 (1992)
13. Handl, J., Knowles, J.: An evolutionary approach to multiobjective clustering. IEEE Transactions on Evolutionary Computation **11**(1), 56–76 (2007)
14. Jain, M., Jain, M., AlSkaif, T., Dev, S.: Which internal validation indices to use while clustering electric load demand profiles? Sustainable Energy, Grids and Networks **32**, 100849 (2022)
15. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)
16. Raschka, S., Mirjalili, V.: Python Machine Learning. Packt Publishing (2015)
17. Rokach, L., Maimon, O.: Clustering methods. In: Clustering Methods Data Mining And Knowledge Discovery Handbook, Maimon, O. And Rokach, L. (eds) (2005)
18. Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. Univ Kans sci bull **38**, 1409–1438 (1958)
19. Sorensen, T.A.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biol. Skar. **5**, 1–34 (1948)
20. Yapiz: Evolutionary clustering and automatic clustering. `https://www.mathworks.com/matlabcentral/fileexchange/52865-evolutionary-clustering-and-automatic-clustering` (2022), retrieved February 2, 2022.