A Collaborative Multi-objective Approach for Clustering Task Based on Distance Measures and Clustering Validity Indices[⋆]

Beatriz Flamia Azevedo[1,2,3] [0000−0002−8527−7409], Ana Maria A. C. Rocha[2] [0000−0001−8679−2886], and Ana I. Pereira[1,2,3] [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] Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Bragança - 5300-253, Portugal. {beatrizflamia, apereira}@ipb.pt, arocha@dps.uminho.pt

Abstract. Clustering algorithm has the task of classifying a set of elements so that the elements within the same group are as similar as possible and, in the same way, that the elements of different groups (clusters) are as different as possible. This paper presents the *Multi-objective Clustering Algorithm* (MCA) combined with the NSGA-II, based on two intra- and three inter-clustering measures, combined 2-to-2, to define the optimal number of clusters and classify the elements among these clusters. As the NSGA-II is a multi-objective algorithm, the results are presented as a Pareto front in terms of the two measures considered in the objective functions. Moreover, a procedure named *Cluster Collaborative Indices Procedure* (CCIP) is proposed, which aims to analyze and compare the Pareto front solutions generated by different criteria (Elbow, Davies-Bouldin, Calinski-Harabasz, CS, and Dumn indices) in a collaborative way. The most appropriate solution is suggested for the decision-maker to support their final choice, considering all solutions provided by the measured combination. The methodology was tested in a benchmark dataset and also in a real dataset, and in both cases, the results were satisfactory to define the optimal number of clusters and to classify the elements of the dataset.

Keywords: clustering validity indices · multi-objective · classification.

[⋆] 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, 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 one of the most widely used methods for unsupervised learning. Its main purpose is to divide the elements of a dataset into groups (clusters) based on the similarities and dissimilarities of the elements. A good clustering algorithm should maintain high similarity within the cluster and higher dissimilarities in distinct clusters. Most current clustering methods have also been proposed for integrating different distance measures to achieve the optimum clustering division. However, the weights for various distance measures are challenging to set [15]. So, a multi-objective optimization algorithm is a suitable strategy for this problem. Besides, in many cases, the estimation of the number of clusters is difficult to predict due to a lack of domain knowledge of the problem, clusters differentiation in terms of shape, size, and density, and when clusters are overlapping in nature [9]. Thus, providing a set of optimal solutions (multi-objective approach) instead of a single one (single-objective approach) is more effective, mainly in problems where human knowledge (decision-maker) is essential.

The advantage of using multi-objective strategies in the clustering task is to combine multiple objectives in parallel. In this way, it is possible to consider different distance measures and cluster quality parameters to provide a more robust and flexible algorithm. Thus, some research deeply explored these advantages in recent years. Kaur et al. [14] explore compactness and connectedness clustering properties through a multi-objective clustering algorithm based on vibrating particle system; Nayak et al. [17] present a multi-objective clustering combined with the Differential Evolution algorithm, based on three objectives related to closeness and separation between the cluster elements and also minimization of the number of the clusters; Liu et al. [15] present two multi-objective clustering approaches based on the combination of multiple distance measures; Dutta et al. [9] proposes a Multi-Objective Genetic Algorithm for automatic clustering, considering numeric and categorical features, that take advantage of the local search ability of k-means with the global search ability of MOGA to find the optimum k, intending to minimize the intra-cluster distance and maximize the inter-cluster distance. All of these presented approaches promise results for classifying elements of different datasets.

The approach proposed in this work explores different clustering measures (two intra- and three inter-clustering measures), combined 2-to-2, to develop a flexible and robust multi-objective clustering algorithm, not dependent on the initial definition of the number of centroids. For this, a Multi-objective Clustering Algorithm (MCA) was developed combined with the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [6], with two intra- and three inter-clustering measures in parallel, minimizing the intra-clustering measure and maximizing the inter-clustering measure. For the six possible combinations, a Pareto front was generated, and the solutions were evaluated by five clustering validity indices (CVIs): Elbow (EI), Davies-Bouldin (BD), Calinski-Haranasz (CH), CS, and Dumn (DI) indices, through the Cluster Collaborative Indices Procedure (CCIP). This evaluation aims to refine the Pareto front solutions and support the decision-makers final choice based on different metrics proposed by each

CVIs, collaboratively. The collaborative algorithm is very helpful in case the decision-maker does not know enough to select one solution from the Pareto front set since the method can suggest the most appropriate solution among the ones that belong to the Pareto front sets.

This paper is organized as follows. After the introduction, Sect. 2 describes the clustering measures, which are divided into intra- and inter-clustering measures. After that, Sect. 3 presents the clustering validity indices (CVIs). Section 4 presents the algorithm developed, the Clustering Multi-objective Algorithm (MCA), and the Cluster Collaborative Index Procedure (CCIP). The results and discussions are presented in Sect. 5. Finally, Sect. 6 presents the conclusion and future steps.

2 Clustering Measures

To classify the elements of the dataset into different groups, it is necessary to establish some measures for computing the distances between elements. The choice of distance measures is fundamental to the algorithm's performance since it strongly influences the clustering results. In this work, different clustering measures are considered to automatically define the optimal number of clusters, minimizing the intra-cluster distance and simultaneously maximizing the inter-cluster distance in a multi-objective approach.

Consider a dataset $X = \{x_1, x_2, ..., x_m\}$, where each observation is a $|d|$ - dimensional real vector. The clustering algorithm consists of partitioned the elements of X into k subsets, it is clusters, in which each cluster set is defined as $C_j = \{x_1^j, x_2^j, ..., x_i^j\}$ with $j = \{1, ..., k\}$, in other words, x_i^j represents an element i that belongs to cluster j and, on the other hand, x_l^t represents another element l that belongs to cluster t. Following, Sect. 2.1 and Sect. 2.2 present the intra- and inter-clustering measures considered, respectively.

2.1 Intra-clustering Measures

Intra-clustering measures refer to the distance among elements of a given cluster. There are many forms to compute the intra-clustering measure. Based on previous studies [3], two of them are explored in this paper, as presented bellow:

- **SMxc**: mean distances between the elements belonging to cluster C_j until its centroids, c_j.
- **FNc**: sum of the furthest neighbor distance of each cluster c_j, where x_i^j and x_l^j belong to the same cluster c_j.

2.2 Inter-clustering Measure

In turn, inter-clustering measures define the distance between elements that belong to different clusters or about the distance between different centroids c_j. In this case, three inter-clustering measures were considered [3]:

- **Mcc**: mean of the distance of all centroids .
- **MFNcc**: mean of the distances of the furthest neighbors among the different clusters, in terms of the number of clusters.
- **MNNcc**: mean of the nearest neighbor distance between elements of the different clusters.

3 Cluster Validity Indices

Cluster Validity Indices (CVIs) define a relation between intra-cluster cohesion and inter-cluster separation to assess the clustering separation quality. A CVI is expected to be able to distinguish between superior and inferior potential solutions, to guarantee the efficiency of the clustering algorithm [13]. The CVI outcome depends only on the partition provided by the clustering algorithm given a specific number of groups [10]. An optimal solution for one specific CVI could not be the optimal solution for another CVI, since each of them has shortcomings and biases [12]. In this way, there are several CVIs available in the literature, as well as several comparative studies between them, as can be seen in [1,10]. For this reason, in this work, it was chosen to use multiple CVIs, through a collaborative strategy, to reduce their shortcomings and biases. Thus, five of them were chosen, the classical ones according to the literature, and they are described below.

3.1 Elbow Index

To use the Elbow index (EI) it is necessary to evaluate the Within-Cluster Sum of Square (WCSS), which means the sum of the Euclidean distance between the elements to their centroids j, for each cluster, given by the Equation (1). Therefore, the WCSS is the sum of all individual $WCSS_j$. When the number of clusters k is less than the optimal number of clusters, WCSS should be high, and when it increases, WCSS will follow an exponential decay. At some point, the decay will become almost linear and WCSS will continue to fall smoothly. The first point that deviates from the exponential curve is considered the elbow, and the associated number of clusters is selected as the optimum. A simplified graphic approximation to find the elbow is to draw a straight line between the WCSS values of the first (with $k = k_{min}$) and last ($k = k_{max}$) cluster solutions and calculate the distance between all the points on the curve and the straight line. Thence, the elbow is the point with the highest distance to the line [7].

$$WCSS_j = \sum_{i=1}^{\#C_j} D(x_i^j, c_j) \tag{1}$$

3.2 Davies-Bouldin Index

The Davies-Bouldin index (DB) [1], 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 $D(c_j, c_t)$. 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 (2),

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

in which $D(x_i^j, c_j)$ is the Euclidean distance between an element x_i^j, that belong to the cluster j, and its centroid c_j. Thus, the DB index is given by Equation (3), where $D(c_j, c_t)$ is the Euclidean distance between the centroid c_j, and the centroid c_t, and the k is the number of clusters. The smallest DB indicates the optimal partition.

$$DB = \frac{1}{k} \sum_{j=1}^{k} \max_{t=1,\ldots k, j \neq t} \left\{ \frac{S(c_j) + S(c_t)}{D(c_j, c_t)} \right\} \tag{3}$$

3.3 Calinski-Harabasz Index

The Calinski-Harabasz (CH) [4] is a ratio-type index where the cohesion is estimated based on the distance from the elements in a cluster to its centroid [1,4]. 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. The BGSS is defined as Equation (4)

$$BGSS = \sum_{j=1}^{k} \#C_j \times D(c_j, \overline{X}) \tag{4}$$

The second step is to calculate the intra-cluster dispersion for each cluster j, also given by the sum all individual within group sums of squares, $WCSS$, as defined in Equation (1). Thus, the CH index is defined as Equation (5):

$$CH = \frac{\#X - k}{k - 1} \times \frac{BGSS}{WCSS} \tag{5}$$

3.4 CS Index

The CS index [5] is a ratio-type index that estimates the cohesion by the cluster diameters and the separation by the nearest neighbor distance. This measure is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The smallest CS, defined by Equation (6) indicates a valid optimal partition [5].

$$CS = \frac{\sum_{j=1}^{k}\{\frac{1}{\#C_j}\sum_{x_i^j \in \#C_j} \max_{x_l^j \in C_j}\{D(x_i^j, x_l^j)\}\}}{\sum_{j=1}^{k}\{\min_{t \in 1:k, t \neq j}\{D(c_j, c_t)\}\}} \tag{6}$$

3.5 Dumn Index

The Dumn index (DI) [8] 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 [8]. So, DI is defined as the rate between the minimum distance between elements of different clusters, it is x_i^j and x_l^t, and the largest distance between elements of the same cluster, it is x_i^j and x_l^j (sometimes called cluster diameter), as defined in Equation (7).

$$DI = \frac{\min_{j,t \in 1:k}\{D(x_i^j, x_l^t)\}}{\max_{j=1:k}\{D(x_i^j, x_l^j)\}} \tag{7}$$

4 Proposed Algorithms

This section presents the *Multi-objective Clustering Algorithm* (MCA) that, together with the NSGA-II, consists of evaluating intra- and inter-clustering measures to define the optimal number of cluster partitions (centroids) and their optimal position, minimizing the intra-cluster distance and maximizing the inter-cluster distance. As we are considering six pairs of measures, the results of the approach are six Pareto fronts, one from each pair of solutions. Furthermore, a procedure denoted as *Cluster Collaborative Indices Procedure* (CCIP) is proposed, which aims to combine and refine the Pareto front solutions using different CVI criteria, in a collaborative way. Thence, the most appropriate solution, according to all CVIs, is selected to support the decision-maker's final choice.

4.1 Multi-objective Clustering Algorithm

To explain the MCA, consider the dataset $X = \{x_1, x_2, ..., x_m\}$ composed of m elements which are intended to partition X into k groups (clusters). As the MCA can automatically define the optimal number of cluster partitions, it is necessary to define the range of possible partitions; it is the minimum and maximum number of centroids k. So, it was defined k_{min} as the minimum number of centroids, and $k_{max} = [\sqrt{m}]$ the maximum number of clusters that the dataset can be partitioned, where k_{max} corresponds to the integer value of the square root of the number of elements in the dataset X.

Next, the MCA randomly generates k_{max} ordered vector belonging to the domain of X, which are the possible candidates for the centroids. For each candidate, a random value ω belonging to $[0, 1]$ is associated, which will be used to select the centroids based on a threshold value γ. The centroids candidates that satisfy the constraint $\omega > \gamma$ advance to the next selection phase.

Following, the Euclidean distance between all elements from X to all centroid candidates k is evaluated. The elements closest to each centroid k define a cluster set C. To avoid small clusters sets, the centroids k that have less than α associated elements, in which $\alpha = [\sqrt{m}]$, are removed from the set of centroids and the elements become part of one remaining centroid, which is the closest one in terms of Euclidean distance of the elements. The remaining centroids are denoted as the centroid of each subset c_k, in which X is partitioned.

After all elements are associated with a centroid c_j, a position must be set at each coordinate to improve the performance of the algorithm. Thus, the coordinates of each centroid assume the coordinates of its cluster barycenter, composed of its elements, x_i^j.

Next, the objective functions values f_h of the problem are calculated, for $h = 2$, where f_1 represents an intra-clustering measure, chosen among the ones presented at Sect. 2.1 and f_2 represents an inter-clustering measure, chosen among the ones presented at Sect. 2.2. Therefore, The NSGA-II algorithm [6] was used to define the set of optimal solutions to the problem, that is, to define finding the Pareto front. By default, the NSGA-II is a minimization algorithm, so the f_2 values are considering negative, respecting the principle of $min\ f_2 = -max\ f_2$ [6].

4.2 Cluster Collaborative Indices Procedure

To evaluate the quality of the solutions of the Pareto fronts generated, the Cluster Collaborative Indices Procedure (CCIP) was developed. In this procedure, each solution of each Pareto front was evaluated by each CVIs criterion. As previously said, an optimal solution for one specific CVI could not be optimal for another CVI [12]. So, after evaluating each solution according to each CVI, the CCIP selects the β best solutions according to each CVIs criterion. Next, the intersection solutions between each Pareto front are evaluated, that is $FS = PF_1 \cap PF_2 \cap PF_3 ... PF_b$, where b is the maximum number of Pareto fronts to be evaluated, and FS defines the set of final optimal solutions composed of the solution provided by the different Pareto fronts. In this way, the solutions defined in FS are among the best β of each pair of clustering measures combination, according to all five CVIs. After that, each CVI indicates its best solution from the remaining set FS to assist the decision-maker in the solution selection. The solution with the most indications is considered the most appropriate to be selected. In case of a tie, the set of solutions indicated is considered the most appropriate for the problem, and it is up to the decision maker to take the final decision.

5 Results and Discussion

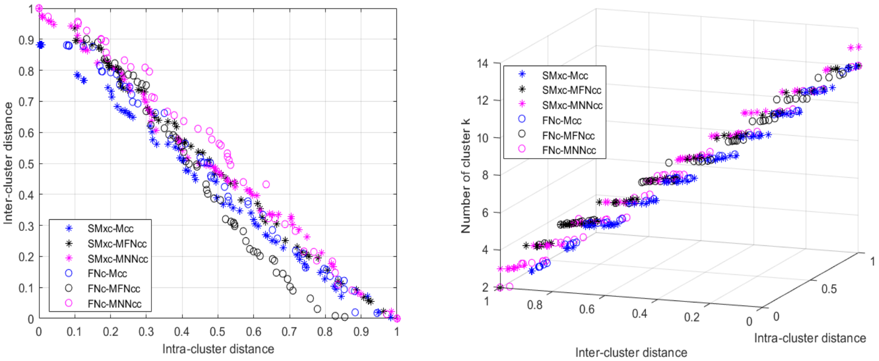
To evaluated the approaches proposed, two datasets are considered. For both dataset, the MCA parameters used were $k_{min} = 2$, $k_{max} = [\sqrt{m}]$, $\gamma = 0.4$. Since MCA is a stochastic algorithm, 10 runs was considered for each measure combination. Regarding NSGA-II parameters, a population equal to 100, maximum generation equal to $200 \times n_f$ were used, where n_f is the number of features, as default [16]. For NSGA-II, the algorithm stops when the geometric mean of the relative change in spread value over 100 generations is less than 10^{-4}, and the final spread is less than the mean spread over the past 100 generations, as defined in `gamultiobj` function [16] documentation.

5.1 Results from Dataset 1

The dataset 1 is a benchmark dataset composed of 300 elements and 2 attributes, available at [11], which indicates 3 clusters as the optimal solution by a single objective approach. Thus, the two intra-clustering measures ($SMxc$ and FNc) are combined with the three inter-clustering measures (Mcc, $MFNcc$, and $MNNcc$), 2-to-2, with the first objective function being an intra-clustering measure, and the second objective function being an inter-clustering measure. This combination results in six Pareto fronts, but with different ranges since they involve measures of sums and means. Thus, to have a fair comparison between the Pareto fronts, they have been normalized. The results of this manipulation are presented in Figure 1a and Figure 1b illustrating the same six Pareto fronts, but also showing the number of clusters k on the z-axis.

Note that, according to the algorithm parameter, the maximum number of k allowed is 17. However, in the model solution, 14 was the maximum number of clusters in a Pareto front, provided by the combination $SMxc$ and $MNNcc$, while for the other combinations, the maximum k is 13. As for the number of solutions, 456 solutions were found, being 91, 78, 97, 69, 61, and 60 solutions, respectively for the combinations $SMxc - Mcc$, $SMxc - MFNcc$, $SMxc - MNNxx$, $FNc - Mcc$, $FNc - MFNcc$, and $FNc - MNNcc$. It is important to highlight that all solutions presented in Figure 1 are optimal for the problem since they all belong to a non-dominated solution set of solutions of their pair of measures.

After that, these solutions were evaluated by the five CVIs indices (EI, DB, CH, CS, and DI). Since the solution is intended to be refined according to the criteria of each CVI, $\beta = 0.75$ was defined, which means that the 75% best solutions according to each CVI are kept, and the intersection set of these remaining solutions is calculated. Figure 2 illustrates the results of the Pareto fronts. Thus, Figure 2a illustrates the Pareto fronts result, and Figure 2b illustrates the Pareto fronts with the number of cluster k indication on the z-axis. In this case, it is possible to note that the maximum k is 9. So, all solutions with k larger than 9 were removed, as well as the other solutions that do not belong to the 75% best solutions for each CVI. Thus, the FS set is composed by 118 solutions, it is, 16 for the combinations $SMxc - Mcc$, 16 for $SMxc - MFNcc$,

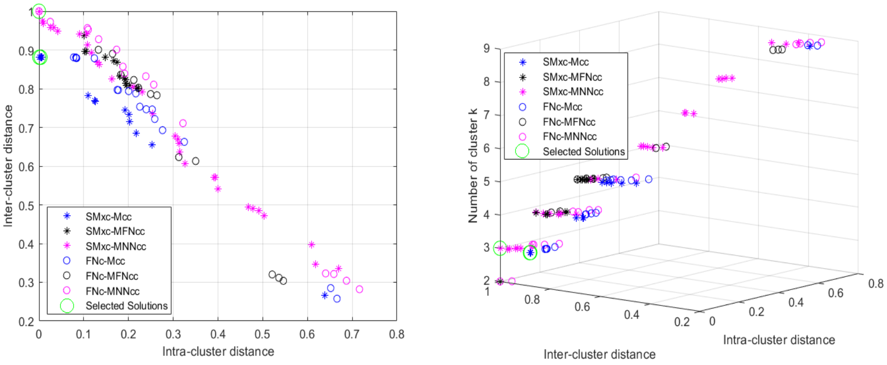


(a) Pareto fronts

(b) Figure 1a with k indication

Fig. 1: Pareto fronts of dataset 1

37 for $SMxc - MNNxx$, 20 for $FNc - Mcc$, 14 for $FNc - MFNcc$, and 15 for $FNc - MNNcc$, which represent a 74% reduction relatively to the initial set of optimal solutions.



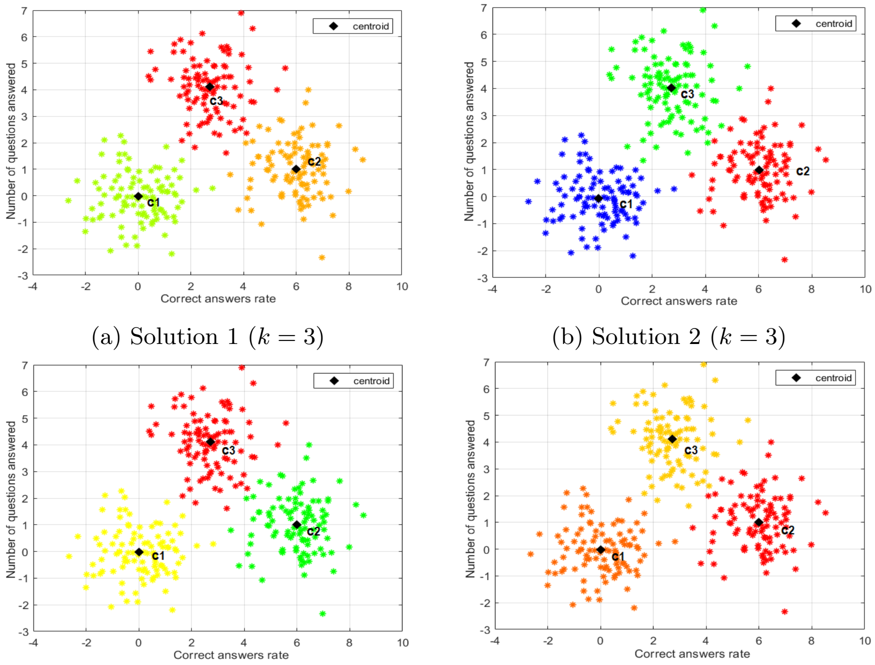
(a) Pareto fronts with $\beta = 0.75$

(b) Figure 2a with k indication

Fig. 2: Pareto fronts of dataset 1, considering $\beta = 0.75$

After that, each CVI identifies its best solution from the remaining set. Thus, the solution with the most indications is considered the most appropriate to be selected. Furthermore, in case of a tie, the set of solutions indicated is also considered the most appropriate for the problem. Considering the remaining solutions, in dataset 1, there was a tie between four solutions provided by indication of the indices: DB, CH, CS. Figure 3 illustrates these four solutions. Although they all divide the dataset into 3 sets, the centroids' position and the distribution of the elements are different.

Solution 1 and 2 are provided by the objective function 1 being the $SMxc$ and the objective function 2 being the Mcc; while solution 3 and 4 are provided by $SMxc$ and $MNNcc$, objective functions 1 and 2, respectively. Thereby, solution 1 centroids are denoted as $c_1 = (0.019, -0.032)$, $c_2 = (5.978, 1.004)$, and $c_3 = (2.713, 4.104)$, that can be analyzed in Figure 3a. Solution 2 centroids are $c_1 = (-0.008, -0.065)$, $c_2 = (6.011, 0.977)$, and $c_3 = (2.719, 4.020)$ - Figure 3b. Solution 3 centroids are $c_1 = (0.019, -0.032)$, $c_2 = (5.978, 1.003)$, $c_3 = (2.713, 4.105)$ - Figure 3c. And, solution 4 centroids are defined as $c_1 = (0.019, -0.032)$, $c_2 = (5.978, 1.004)$, and $c_3 = (2.713, 4.105)$ - Figure 3d. Thus, according to the results, there is no doubt that the most appropriate number of k is 3, which goes to the solution of [11]. Thence, it is only up to the decision-maker to choose (if necessary) the distribution of the elements for the problem or just select one of the four solutions, that are approximately equal solutions.



(a) Solution 1 ($k = 3$)

(b) Solution 2 ($k = 3$)

(c) Solution 3 ($k = 3$)

(d) Solution 4 ($k = 3$)

Fig. 3: Final Pareto front solutions (dataset 1)

5.2 Results from Dataset 2

To test the approach on a real case study, the methodology previously presented for dataset 1, was also applied for the dataset 2. Dataset 2 is a real case study composed of 291 elements and 2 instances, and it is provided by the MathE project [2]. The MathE project aims to provide any student all over the world with an online platform to help them to learn college mathematics and also support students who want to deepen their knowledge of a multitude of mathematical topics at their own pace. More details about the MathE project are described in [2], and can also be found on the platform Website (mathe.pixel-online.org). One of the particularities of the MathE platform is the *Student's Assessment* section, which is composed of multiple-choice questions for the students to train and practice their skills. The answers provided by each student over the 3 years that the platform has been online define the dataset 2. Therefore, each dataset element refers to one student who used the Student Assessment section. And the first instance represents the rate of the correct answer (x-axis) provided by the student's history, and the second instance represents the number of questions answered by this student (y-axis) while MathE user. To support the result analysis, the y-axis, which initially varies from 1 to 42 (number of questions answered), has been normalized by range; it is between 0 to 1.

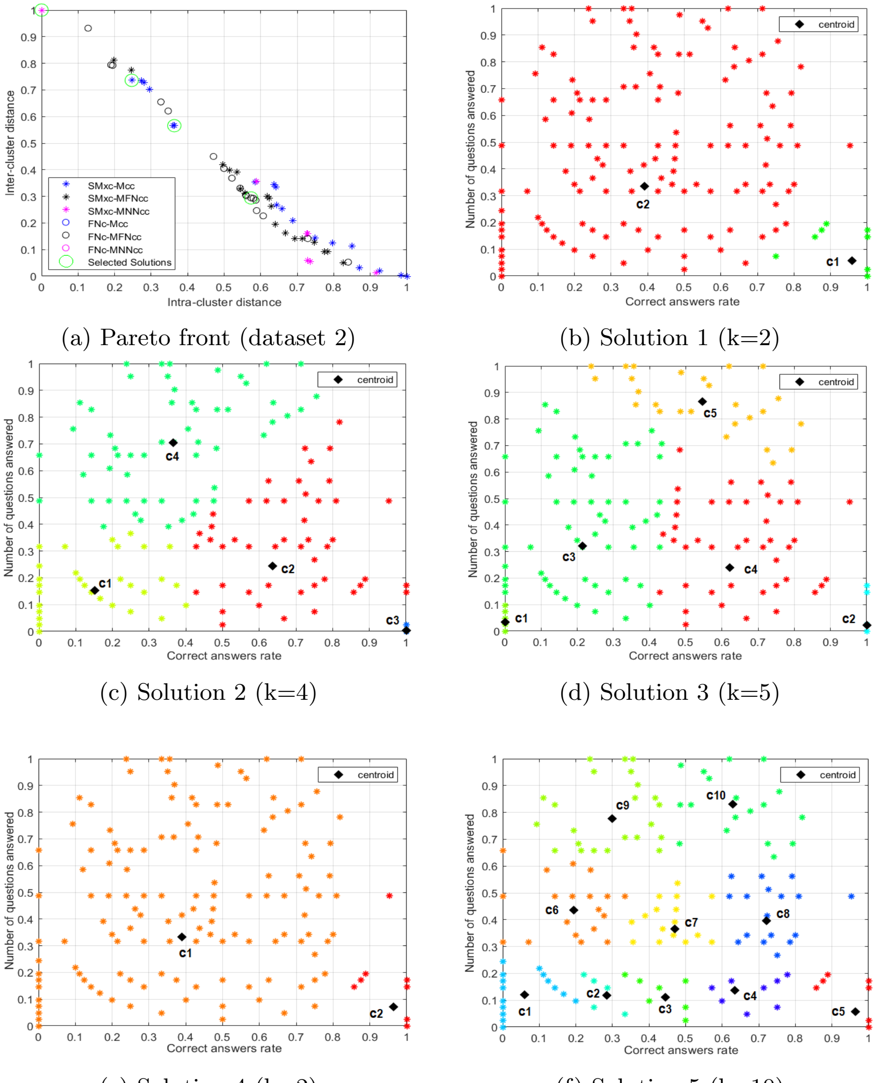
Preliminary studies involving cluster classification and MathE students' data, but using a single objective approach, did not show satisfactory results [2]; that is, the patterns extracted did not provide the necessary information to be used by the project. This is because the single objective algorithm only provides a single solution, which, although optimal, is not relevant to the decision-maker's request. For this reason, the dataset 2 is an excellent example to be analyzed with the proposed approach since the choice of the optimal solution is strongly dependent on the sensitivity of the decision-maker.

The methodology described for dataset 1 is applied to dataset 2, i.e., six Pareto fronts were generated and normalized, the 75% best solutions for each CVIs were considered, and the intersection set of these solutions was evaluated. In this case, the initial set of all Pareto fronts is composed of 312 solutions (50 of $SMxc - Mcc$, 46 of $SMxc - MFNcc$, 54 of $SMxc - MNNxx$, 60 of $FNc - Mcc$, 48 of $FNc - MFNcc$, and 54 of $FNc - MNNcc$), and after the refinement, the final set FS is composed by 64 solutions (21 of $SMxc - Mcc$, 18 of $SMxc - MFNcc$, 7 of $SMxc - MNNxx$, and 18 of $FNc - MNNcc$). An 80% reduction in the number of optimal solutions is verified and the result of this approach is presented in Figure 4a. After that, each CVI indicates its most appropriated solution. For dataset 2 each CVI indicated one different solution, as presented in Figures 4b- 4f, in which solution 1 was indicated by EI, solution 2 by DB, solution 3 by CH, solution 4 by CS, and solution 5 by DI. Thus, solutions 1, 2, and 3 were provided by objective function 1 equal to $SMxc$ and objective function 2 equal to Mcc. Whereas, solutions 4 were given by objective function 1 equal to $SMxc$ and objective function 2 equal to $MNNcc$. And, solutions 5 were resulted by objective function 1 equal to FNc and objective function 2 equal to $MFNcc$. Thereby, the centroids of solution 1 are $c_1 = (0.958, 0.058)$

and $c_2 = (0.391, 0.335)$, in Figure 4b. The centroids of solution 2 are $c_1 = (0.152, 0.153)$, $c_2 = (0.636, 0.245)$, $c_3 = (1, 0.004)$, and $c_4 = (0.365, 0.705)$, as depicted in Figure 4c. The centroids of solution 3 are $c_1 = (0.000, 0.033)$, $c_2 = (1, 0.023)$, $c_3 = (0.215, 0.321)$, $c_4 = (0.621, 0.239)$, and $c_5 = (0.546, 0.866)$, as can be seen in Figure 4d. The 4 solution centroids are $c_1 = (0.390, 0.333)$, and $c_2 = (0.964, 0.071)$ - Figure 4e. The centroids of solution 5 are $c_1 = (0.060, 0.120)$, $c_2 = (0.284, 0.117)$, $c_3 = (0.445, 0.111)$, $c_4 = (0.635, 0.137)$, $c_5 = (0.965, 0.057)$, $c_6 = (0.194, 0.435)$, $c_7 = (0.469, 0.366)$, $c_8 = (0.720, 0.396)$, $c_9 = (0.298, 0.777)$, and $c_{10} = (0.629, 0.831)$, in Figure 4f.

Knowing the profile of students enrolled in the MathE platform, it is known that there is a diversity of students with different backgrounds (country, age, course and university year attending, and level of difficulty in Mathematics, among others). Therefore, a division into a few groups is not a significant result for the project, given the diversity of the public, especially in terms of performance in mathematical disciplines, as already explored in previous works. Thus, considering the previous information and interest of the MathE Project, solution 5, in Figure 4f, is chosen as the most appropriate real one.

In solution 5, the dataset was divided into 10 clusters. In terms of the number of questions answered, clusters 1 to 5 are composed of students who answer a few questions. In contrast, clusters 6, 7, and 8 comprise students who answer a larger number of questions than the previously mentioned groups. Finally, clusters 9 and 10 are made up of students who answered the most quantity of questions on the platform. In terms of performance (correct answers rate), considering clusters 1 to 5, the students' performance increases gradually for cluster 1 to cluster 5, so in cluster 1 almost all students have a success rate equal to 0, while in cluster 5 almost all students had 1. Here it is important to point out that dataset 2 is composed of multiple equal entries (student with an equal number of questions answered and equal performance), which overlap on the graph; for this reason, cluster 5, although it seems to be composed by few students, actually includes 22 students, with 17 having 1 question answered and 1 correct answer. In clusters 6, 7, and 8 the students used the platform more than in the previous groups. In this case, the students of cluster 6 performed less than 0.35, whereas the students of cluster 7 performed between 0.35 and 0.6, and the students who performed higher than 0.6, are in cluster 8. In clusters 9 and 10, the students answered more questions. Regarding their performance, in cluster 9 they have a performance lower than 0.45, while in cluster 10 the student rate performance is higher than 0.45. In this way, the division provided by solution 5 can be used to extract valuable characteristics about the student's performance according to the group to which they belong.



(a) Pareto front (dataset 2)

(b) Solution 1 (k=2)

(c) Solution 2 (k=4)

(d) Solution 3 (k=5)

(e) Solution 4 (k=2)

(f) Solution 5 (k=10)

Fig. 4: Pareto front and solutions (dataset 2), considering $\beta = 0.75$

6 Conclusion and Future Works

The advantage of using multi-objective strategies in the clustering task is to combine multiple objectives in parallel, such as different distance measures. This paper explored clustering measures to develop the Multi-objective Clustering Algorithm. The results of MCA consist of a set of Pareto front solutions, provided by the pair of measures, that were considered as the objective functions of a bi-objective optimization problem. The problem aimed to minimize the intra-clustering measure and maximize the inter-clustering measure using the NGSA-II. In this case, the objective function 1 refers to the intra-clustering measures, which could be the measures $SMxc$ or FNc; and the objective function 2 refers to the inter-clustering measures, which could be any one of the measures Mcc, $MFNcc$, or $MNNcc$. Besides, a procedure denoted as Cluster Collaborative Indices Procedure was proposed, aiming to compare and refine the Pareto front solutions generated by the MCA and NSGA-II, using different criteria provided by five CVIs: Elbow (EI), Davies-Bouldin (BD), Calinski-Haranasz (CH), CS, and Dumn (DI) indices. Thus, the optimal β solutions were selected according to each CVI, and the worst $(1-\beta)$ solutions of each CVI are removed. The intersection set between each CVI β solution is calculated; finally, each CVI indicates its most appropriate solution of the intersection set. The solution with more indications is suggested to the decision-maker as the most appropriate one.

By the range and variability of each Pareto front generated, it is possible to perceive the impact of combining different measures to solve a problem. Analyzing the results of dataset 1, by Figure 1b, it is evident that only the combination $SMxc - MFNcc$ provided solutions with $k = 14$, whereas the combination $FNc - MFNcc$ does not have solutions with k less than 4. In this way, if only one pair of measures were considered, the final solution was restricted to the optimum provided by the pair of measures combination. So, considering the results of the six Pareto fronts, the final solution is enriched by the solution provided by different measures. As already mentioned by [12], an optimal solution for one specific CVI could not be the optimal solution for another CVI due to their metrics. Considering this, choosing the most appropriate CVI for the problem is not a simple task. The intersection strategy serves to refine the solutions and ensure that all the remaining are the most appropriate β for each of the CVI, as it is very hard to achieve a solution that is the best for all CVI.

The indication of the best CVI solution is useful to help the decision-maker since even after selecting the most appropriate optimal solutions, there are still many options left, and in certain cases, the decision-maker does not have enough information about the data to quickly determine, among the set of optimal solutions, the one that most represent the problem. According to [11], considering a single objective strategy, the optimal solution for dataset 1 is 3 clusters, it is $k = 3$. As shown in Figure 3, all the solutions indicated from dataset 1 have $k = 3$, demonstrating the effectiveness of the proposed method in a benchmark problem.

In the case of dataset 2, the data distribution is more complex than in dataset 1 [9], since the multiple points and clusters overlap, are not rounded shape, and the elements are not as well separated as the dataset 1. Thus, for dataset 2, which describes a real problem, the multi-objective strategy is much more effective than the single one since in the multi-objective, it is possible to compare and choose among a set of optimal solutions, the one that goes from meeting the patterns that the decision-maker wants to extract from the dataset. For the dataset 2, the decision maker's knowledge is of great value in defining the solution to be used.

Then, the proposed method is an asset in situations where the single objective approach is insufficient.

In the future, it is expected to explore more deeply the intra- and inter-clustering measures in multiple objective functions, as well as cluster splitting and merging strategies to improve the quality of cluster partitioning.

References

1. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recog. **46**(1), 243–256 (2013)
2. 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: Book of 16th Learning and Intelligent Optimization Conference - LION 2022. pp. 319–333. Milos - Greece (2022)
3. Azevedo, B.F., Rocha, A.M.A.C., Pereira, A.I.: A multi-objective clustering approach based on different clustering measures combinations. Submitted to Computational & Applied Mathematics
4. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974)
5. Chou, C.H., Su, M.C., Lai, E.: A new cluster validity measure and its application to image compression. Pattern Analysis and Applications **7**, 205–220 (2004)
6. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Transactions on Evolutionary Computation **6**(2), 182–197 (2002)
7. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Novel clustering selection criterion for fast binary key speaker diarization. In: 16th Annual Conference of the International Speech Communication Association (NTERSPEECH 2015) (2015)
8. 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)
9. 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)
10. Gurrutxaga, I., Muguerza, J., Arbelaitz, O., Pérez, J.M., Martín, J.I.: Towards a standard methodology to evaluate internal cluster validity indices. Pattern Recognition Letters **32**(3), 505–515 (2011)
11. Heris, M.K.: Evolutionary data clustering in matlab. https://yarpiz.com/64/ypml101-evolutionary-clustering (2015)
12. 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)
13. José-García, A., Gómez-Flores, W.: A survey of cluster validity indices for automatic data clustering using differential evolution. In: Proceedings of the Genetic and Evolutionary Computation Conference. p. 314–322 (2021)
14. Kaur, A., Kumar, Y.: A multi-objective vibrating particle system algorithm for data clustering. Pattern Anal. Appl. **25**(1), 209–239 (2022)
15. Liu, C., Liu, J., Peng, D., Wu, C.: A general multiobjective clustering approach based on multiple distance measures. IEEE Access **6**, 41706–41719 (2018)
16. MATLAB: Mathworks inc. www.mathworks.com/products/matlab.html (2019a)
17. Nayak, S.K., Rout, P.K., Jagadev, A.K.: Multi-objective clustering: a kernel based approach using differential evolution. Connection Science **31**(3), 294–321 (2019)