Evaluating Student Behaviour on the MathE Platform - Clustering Algorithms Approaches *

Beatriz Flamia Azevedo et al[1,2][0000−0002−8527−7409], Ana Maria A. C. Rocha[2][0000−0001−8679−2886], Florbela P. Fernandes[1][0000−0001−9542−4460], Maria F. Pacheco[1,3][0000−0001−7915−0391], and Ana I. Pereira[1,2][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, University of Minho, Campus Azurém, Guimarães - 4800-058, Portugal
[3] Center for Research & Development in Mathematics and Applications CIDMA, University of Aveiro, Aveiro, Portugal
{beatrizflamia, fflor, pacheco, apereira}@ipb.pt, arocha@dps.uminho.pt

Abstract. The MathE platform is an online educational platform that aims to help students who struggle to learn college mathematics as well as students who wish to deepen their knowledge on subjects that rely on a strong mathematical background, at their own pace. The MathE platform is currently being used by a significant number of users, from all over the world, as a tool to support and engage students, ensuring new and creative ways to encourage them to improve their mathematical skills. This paper is addressed to evaluate the students' performance on the Linear Algebra topic, which is a specific topic of the MathE platform. In order to achieve this goal, four clustering algorithms were considered; three of them based on different bio-inspired techniques and the k-means algorithm. 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 take the risk of answering advanced questions, they make even more mistakes, which causes them to return to the basic level questions. Considering these results, it is now necessary to carry out an in-depth study to reorganize the available questions according to other levels of difficulty, and not just between basic and advanced levels as it is.

Keywords: automatic clustering algorithms · optimization · bio-inspired methods · e-learning technology.

1 Introduction

In an era where the Internet and digital resources, in general, are forcing all teaching system levels to reinvent themselves, it becomes necessary and urgent

* This work has been supported by FCT Fundação para a Ciência e Tecnologia within the R&D Units Project Scope UIDB/00319/2020 and UIDB/05757/2020. Beatriz Flamia Azevedo is supported by FCT Grant Reference SFRH/BD/07427/2021.

to implement changes in teaching and learning processes [17]. Moreover, the COVID-19 pandemic showed how much investment in technological resources and literacy is still necessary in order to allow the strengthening of current educational systems and activities, contributing to increase the students' and teachers' interest in the subjects they are involved in. One way to do this is by applying digital educational technologies such as e-learning platforms.

In particular, Mathematics is considered a fundamental area for the construction of a sustainable knowledge economy, one of the great societal challenges of our time [17, 4]. However, this is one subject that students report most problems in learning and, therefore, it is essential to invest in different and engaging ways of teaching and learning mathematics. Today's students demand that their educational environments integrate the digital tools of the twenty-first century, adapting to their modern way of life and, in this context, the MathE learning environment can offer a valuable contribution to improve the students' confidence in their ability to learn mathematics.

MathE (mathe.pixel-online.org) is an e-learning platform where students from all over the world have free access to resources such as videos, exercises, training tests, and pedagogical materials covering several areas of mathematics taught in higher education courses. The MathE project offers an online tool for autonomous learning, available 24 hours per day, 7 days per week, where students can learn mathematics in an engaging way, more varied and more in line with the dynamics of the current generation of students than the traditional methods. MathE's purpose is to provide students and teachers with a new perspective on mathematical teaching and learning, relying on digital interactive technologies that enable autonomous study [4]. At its current stage, the platform is organized into three main sections, *Student's Assessment*, *MathE Library* and *Community of Practice*, in which fifteen mathematics topics are covered, among the ones that are in the classic core of graduate courses. A more detailed description of the sections and the covered topics can be found at [5].

In particular, the *Student's Assessment* section is composed of multiple-choice questions divided into topics, with two levels of difficulty — basic and advanced — among which the students can make their choice. The students can train and practice their skills in the *Self Need Assessment* (SNA) subsection. This subsection aims to provide the student with a self training assessment to test whether a certain topic that he/she has enrolled in is already properly understood. If a student or a teacher believe that the understanding of a given subject needs to be deepened, the student has the possibility of answering another training assessment to measure his/her degree of confidence in order to perform a final assessment. Each training assessment is be randomly generated from an assessments database composed of questions and their corresponding answers. In this way, the same student is able to answer different training assessments on the same topic. After the student submits a self-assessment test, the corresponding grade automatically appears, allowing self-assessment.

The MathE Platform is being improved, so that it becomes even more interactive and gains intelligence for decision making. In this way, it is expected

that in the near future the questions will be addressed to students in an autonomous way instead of in a randomized manner, as it currently is. One of the first necessary steps to achieve this is to recognize patterns in the data obtained so far. Thus, this work aims to evaluate the student's behavior when answering questions under the Linear Algebra topic of the SNA. Considering the obtained results it is expected to obtain information about the student's performance, that is, if they are getting the right answers or the wrong ones.

Currently, there are 99 teachers and 1161 students from different nationalities enrolled in the platform: Portuguese, Brazilian, Turk, Tunisian, Greek, German, Kazakh, Italian, Russian, Lithuanian, Irish, Spanish, Dutch, and Romanian. In this work, the performance of students using the Linear Algebra topic in the SNA section of the MathE platform will be evaluated. Linear Algebra is the most consulted and answered topic of the platform; this fact is not surprising, considering that Linear Algebra is a subject present in almost all curricula of higher education courses that include mathematics. For this reason, it was the topic chosen for the analysis herein described.

To perform the current research, the data collected over 3 years from students from different countries was analyzed by different clustering techniques in order to investigate the similarities and dissimilarities in the profiles of different groups of students in the topic Linear Algebra.

This paper is organized as follows: after the introduction, Section 2 presents an overview of clustering algorithms and also presents some recent work of bio-inspired clustering techniques. Section 3 introduces the clustering algorithms that will be applied in this work. The database composed of MathE student's performance in the MathE Self Need Assessment is described in Section 4. The results are presented and discussed in Section 5. Finally, the main conclusions and the future paths are described in Section 6.

2 An Overview on Clustering Algorithms and Related Works

Clustering is one of the most widely used methods for unsupervised learning and it is very useful in engineering, health sciences, humanities, economics, education, and in many other areas of knowledge that involve unlabeled datasets, i.e., sets of data where there is no defined association between input and output. Thus, clustering algorithms consist of performing the task of grouping a set of elements with similarities in the same group and dissimilarities in other groups [20].

A crucial step in clustering is to assess the member's proximity that composes a dataset and to partition the dataset into groups, considering the similarity and dissimilarity between a pair of elements. The partitioning method is one of the most common strategies used in clustering algorithms. This method provides a dataset partition into a pre-determined number of clusters, not known a priori. Each cluster is represented by its centroid vector, and the clustering process is carried out in an effort to iteractively optimize a criterion function and, at each

execution step, all centroids are updated in an attempt to improve the quality of the final solution [16].

However, partitioning methods are known for their sensitivity to the initial position of the centroid, which may lead to weak solutions, getting stuck at the local optimum if the algorithm starts in a poor region of the problem space [16]. Moreover, the partitioning clustering algorithm heavily depends on the initial values of the cluster centers [8], which define the number of clustering partitions, as it is the number of groups that the dataset will be divided into.

Aiming to overcome these difficulties, the automatic clustering strategies that combine clustering and optimization techniques have helped to surpass these challenges, offering at the same time several improvements in clustering methods. The automatic clustering process consists of solving an optimization problem, aiming to minimize the similarity within a cluster and maximize the dissimilarity between clusters. Thus, most metaheuristic approaches are judged to fit well in the context of the new clustering paradigm [11].

In this context, several studies suggest using nature-inspired metaheuristics to select the optimal number of clusters and find a solution that maximizes the separation between different clusters and minimizes the distance between data points in the same cluster [18]. Eesa and Orman [8] present a bio-inspired Cuttlefish Algorithm (CFA) combined with the k-means algorithm for searching the best cluster centers that can minimize the clustering metrics and avoid getting stuck in local optima. Likewise, Singh [21] suggests using the Whale Optimization Algorithm (WOA) to improve the cluster exploration mechanism and solve the problem of local entrapment. Nemmich et al. [14] use Artificial Bees Colony Algorithm with a Memory Scheme to improve the k-means performance. So, in the approach presented in [14], a simple memory scheme is introduced to prevent visiting sites which are close to previously visited sites and to avoid visiting sites with the same fitness or worse. All of the enumerated approaches were tested on several benchmark datasets as well as, sometimes, on real-life problems, and the authors considered various statistical tests to justify the effectiveness of combining clustering algorithms and metaheuristics.

Nguyen and Kuo [15] present an automatic fuzzy clustering using a non-dominated sorting particle swarm optimization algorithm for categorical data. The method can identify the optimal number of clusters based on two objective functions that minimize the global compactness and fuzzy separation representing intra-cluster and inter-cluster distances. In its turn, [10] proposes a metaheuristic-based Possibilistic Multivariate Fuzzy Weighted c-means Algorithm (PMFWCM) for clustering mixed data (numerical and categorical). In this case, three metaheuristics, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Sine Cosine algorithm (SCA) are used in different combinations with the PMFWCM for cluster analysis. Both authors claimed that the proposed algorithms work efficiently and determine the optimal number of cluster centers.

Another interesting approach is presented by Atabay et al. [2] which propose a clustering algorithm that integrates PSO and k-means algorithms. The sensi-

tivity of the k-means algorithm to the initial choice of the centroids is solved by PSO integration. On the other hand, the ability to rapidly converge by transitioning the center of a cluster from the previous location to the average location of points belonging to that cluster in each iteration is used to accelerate convergence and improve the result of the PSO algorithm.

Considering what was described in the literature review, several approaches can be combined between bio-inspired optimization and clustering techniques, allowing to mitigate or eliminate some of the difficulties encountered by the methods using hybrid techniques. In this work, three bio-inspired metaheuristic approaches are considered, Genetic Algorithm (GA) [22], Particle Swarm optimization [9], and Differential Evolution (DE) [23], in order to find the optimum number of clusters to assess student performance from the MathE dataset. Besides, the results will be compared with k-means clustering.

3 Clustering Approaches

The cluster separation measure incorporates the fundamental features of some of the well-accepted similarity measures often applied to the cluster analysis problem and also satisfies certain heuristic criteria [1]. In this work, the Davies–Bouldin index (DB) [7] will be used as a clustering measure, that will define the number of cluster centroids, which is the number of groups that the dataset will be divided into.

3.1 Davies–Bouldin index

Davies–Bouldin index (DB) is based on a ratio of intra-cluster and inter-cluster distances. It is used to validate cluster quality and also to determine the optimal number of clusters. Consider that cluster C have members $X_1, X_2, ..., X_m$. The goal is to define a general cluster separation measure, S_i and M_{ij}, which allows computing the average similarity of each cluster with its most similar cluster. The lower the average similarity, the better the clusters are separated and the better the clustering results. To better explain how to get the Davies-Bouldin index, four steps are considered [7].

In the first step, it is necessary to evaluate the average distance between each observation within the cluster and its centroid, that is the dispersion parameter S_i, also know as intra-cluster distance, given by Equation (1),

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^q \right\}^{\frac{1}{q}} \tag{1}$$

where, for a particular cluster i, T_i is the number of vectors (observations), A_i is its centroid and X_j is the jth (observation) vector.

The second step aims to evaluate the distance between the centroids A_i and A_j, given by Equation (2), which is also known as inter-cluster distance. In

this case, a_{ki} is the kth component of the n-dimensional vector a_i, which is the centroid of cluster i, and N is the total number of clusters. It is worth mentioning that M_{ij} is the Minkowski metric of the centroids which characterize clusters i and j and $p = 2$ means the Euclidean distance.

$$M_{ij} = \left\{ \sum_{k=1}^{N} |a_k i - a_k j|^p \} \right\}^{\frac{1}{p}} = ||A_i - A_j||_p \tag{2}$$

In the third step, the similarity between clusters, R_{ij}, is computed as the sum of two intra-cluster dispersions divided by the separation measure, given by Equation (3), that is the within-to-between cluster distance ratio for the ith and jth clusters.

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{3}$$

Finally, the last step calculates the DB index, Equation (4), that is, the average of the similarity measure of each cluster with the cluster most similar to it. R_i is the maximum of R_{ij} $i \neq j$, so, the maximum value of R_{ij} represents the worst-case within-to-between cluster ratio for cluster i. Thus, the optimal clustering solution has the smallest Davies-Bouldin index value.

$$DB = \frac{1}{N} \sum_{i=1}^{N} R_i \tag{4}$$

Considering the definition of the DB index, a minimization problem can be defined, whose objective function is the DB index value. Thus, metaheuristics can be used in order to solve this problem as an evolutionary bio-inspired algorithm.

3.2 Evolutionary Bio-inspired clustering algorithms

The algorithms in the class of evolutionary computation start by randomly generating a set (population) of potential solutions. The population is represented by individuals arranged in the search space, which is the space where each variable can have values (some examples are $\mathbb{Z}^n$, $\mathbb{R}^n$, $\{0, 1\}^n$, ...). The search space is delimited by the domain of the objective function, which ensures that all individuals are feasible solutions for the problem [22]. By iteratively applying the genetic operators like selection, crossover, and mutation (the most common ones), the population is being modified to obtain new feasible solutions. This process stochastically discards poor solutions and evolves more fit (better) solutions [6]. Due to the nature of these operators, which are based on Darwin's evolution principles (in which the most adapted individuals of a given population survive whereas the less adapted die to be replaced by their offspring [6, 22]), it is expected that the evolved solutions will become better generation by generation (iteration). Like any iterative process, the evolutionary algorithms require a stopping criterion to stop the search [22]. Some examples of stopping criteria are described in [3].

In this work, three bio-inspired evolutionary algorithms are used. Genetic Algorithms (GA) [22], which is based on the Darwinian principle of survival of the fittest and encoding of individuals; Differential Evolution (DE) [23], which are inspired by the theory of evolution using natural selection; and Particle Swarm Optimization (PSO) [9] that is an evolutionary algorithm, based on the behavior of birds flocking, or fish schooling. Figure 1 shows the GA, DE and PSO flowcharts.

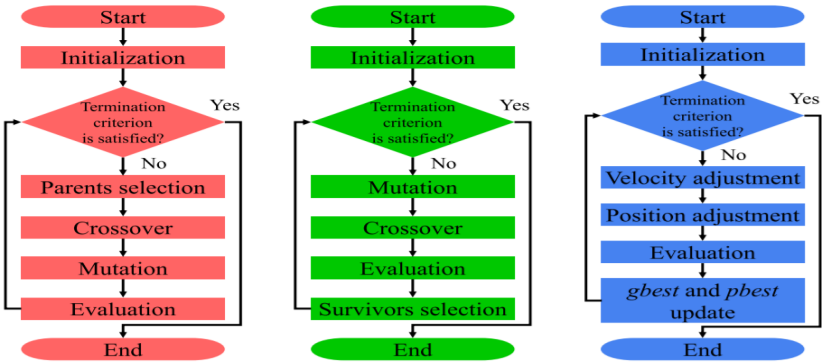

Fig. 1: The GA, DE and PSO flowcharts. Adapted from [13]

The main difference between the variants of the so-called automatic algorithms that will be used in this paper is the optimization process to define the DB-index, since each one of them employs a different bio-inspired optimization algorithm, that is GA, DE or PSO.

3.3 K-means clustering algorithm

The k-means partitioning clustering algorithm is one of the most well-known clustering algorithms, which requires a priori the definition of the number of clusters, being an example of an algorithm that is dependent on the initial solution, as mentioned in Section 2.

The k-means algorithm consists of trying to separate samples into groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (WSS). As k-means is not an automatic clustering algorithm, it requires the definition of the initial parameter k, that represents the number of clusters division. The value of k can be specified by different techniques, such as Silhouette method, Davies-Boulding index, or Calinski Harabasz method [19]. Once this value is established, the k-means algorithm divides a set of X samples $X_1, X_2, ..., X_m$ into k disjoint clusters C, each described by the mean of the samples in the cluster, μ_i, also denoted as cluster "centroids". In this way, the k-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion, presented in Equation (5) [1].

$$WSS = \sum_{i=0}^{m} \min ||X_j - A_i||^2, \text{ in which } \mu_i \in C \tag{5}$$

From these centers, a clustering is defined, grouping data points according to the center to which each point is assigned.

4 Dataset

This study is focused on the analysis of the performance of a set of students on the MathE Student's Assessment section. The data collected and the performed analysis take into consideration information provided by 134 students from different countries who are active and consistent users of the Linear Algebra topic of the Student's Assessment section. These students regularly answer and submit self-assessment tests to support their study and validate their progress on this topic. As was previously mentioned, Linear Algebra is the most accessed topic of the MathE platform, so a considerable amount of basic and advanced questions have been answered. For this reason, this topic was chosen to be analyzed through clustering algorithms.

In order to analyze the students' profile through clustering, the number of questions answered correctly and incorrectly for each student were evaluated, according to the basic and advanced levels. Then, the outlier students were identified through the Box plot method, and these students were removed out of the data set, leaving the information of 99 students for the analysis.

As previously mentioned, the questions available in MathE were divided into two levels of difficulty, basic and advanced. Hence, when a student selects a topic, he/she must also decide the difficult level of questions he/she wants to answer. After that the platform will provide a set of 7 random questions, available in the platform database, that belong from the chosen topic and level of difficult.

Over the 3 years of the platform's availability, 199 different questions were used, out of the 211 available in the platform's linear algebra database (142 basic and 69 advanced), being equal to 3696 the sum of times the available questions were used. Table 1 shows the number of correct and incorrect answers according to the question level. As can be seen, from the 3696 questions answered, 2919 were from the basic level and 777 from the advanced level, making a total of 1741 and 1955 correct and incorrect answers, respectively.

Table 1: Number of question 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**

Table 2 presents the descriptive measures of the considered variables. The *Answers* column refers to the total number of basic or advanced questions that were answered correctly or incorrectly; *Min* and *Max* are the minimum and maximum values obtained in each variable; the column *No. Students* presents the number of students who answered a question correctly or incorrectly at basic or advanced levels. That is, out of the 99 students evaluated, 87 answered at least one basic question correctly and 88 answered at least one basic question incorrectly. On the other hand, only 26 correctly answered at least one advanced question and 23 incorrectly answered at least one advanced question.

Table 2: Descriptive Measures

Variable	Answers	Min	Max	No. Students
Correct Basic	1386	0	25	87
Incorrect Basic	1533	0	31	88
Correct Advanced	355	0	7	26
Incorrect Advanced	422	0	7	23

5 Results and Discussion

The MathE platform has the mission to offer a dynamic and compelling way of teaching and learning mathematics, relying on interactive digital technologies that enable autonomous study [5]. This work focuses on investigating students features, using clustering algorithm in order to recognize patterns in the students platform user's. In the future, these patterns will serve as a guidance to provide intelligence to the platform, making it capable of addressing questions in a personalized way according to each student's profile.

The information of the 99 students who used the Linear Algebra topic were considered in this analysis. The results obtained for the Linear Algebra topic - as was previously mentioned, the most widely chosen - can be inferred for the other less used topics of the platform.

Figure 2 shows the number of questions answered by each one of the 99 considered students, grouped by answered question level, ([a]-basic question and [b]-advanced questions). As already shown in Table 2 and better illustrated in Figure 2, the range of answered basic questions varies from 0 to 35, while the advanced ones vary between 0 and 7. Hence, the figure offers a better perception of the profile of each individual student. It can be clearly seen that the students choose to answer more basic questions than advanced ones. However, even answering more basic than advanced questions, they end up making too many mistakes.

When a student selects a topic and a level on the Self-Need Assessment section, the MathE platform system provides the student with a subset of 7

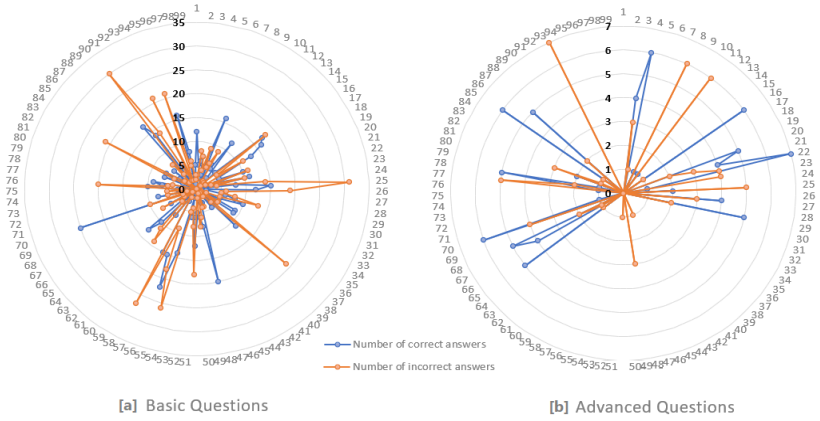

Fig. 2: Student's Performance on basic and advanced questions

questions randomly generated from the assessments database of the selected topic. Thus, when it sums up the number of questions answered by each student at each level, it is possible to evaluate, on average, how many tests these students used to answer. Thus, evaluating by question difficulty level, it can be deduced from Figure 2(a) that the center of the figure, that is, with a range $[0, 10]$, is comprised of students who answer 1 or 2 basic tests, which represent most of the students; within the range $[11, 20]$, are the students answered more than 2 tests. Finally, above the range 20, are the students who answered at least 3 or more tests. Concerning the advanced questions (see Figure 2(b)), we can say that students answer at most 2 tests of 7 advanced questions and do not return to this level afterwards.

Aiming to group the different profiles of students and analyzing the similarities and dissimilarities between students groups, the dataset was evaluated by clustering algorithms. Therefore, three automatic clustering algorithms were used to define the optimum number of clusters and establish their optimum position. Thence, three bio-inspired optimization strategies were considered, namely GA, PSO, and DE. More details about the algorithms' codification can be consulted at [24]. Moreover, the results of these three approaches were compared with k-means algorithm, which is an example of a non-automatic clustering algorithm [1].

For all bio-inspired algorithms, the common parameters used were: maximum number of clusters equal to 10; initial population equal to 100, maximum number of iterations equal to 250, which was also the stoppage criterion considered. For the GA, a rate of 0.8 was considered for selection and crossover, and 0.3 for a mutation. On the other hand, for PSO, the chosen rates were: global learning coefficient equal to 2, personal learning coefficient equal to 1.5, inertia weight equal to 1 and inertia weight damping equal to 0.99. Finally, for DE, the rates are equal to 0.2 for crossover and the scaling bound factor varies between $[0.2, 0.8]$.

The results were obtained using an Intel(R) i5(R) CPU @1.60 GHz with 8 GB of RAM using Matlab software [12].

In order to perform the clustering analysis, 4 variables were defined, as presented below. Each of them describes the number of questions answered by a student according to the question 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

All clustering algorithms considered the four variables at the same time. The number of centroids and their positions defined by each algorithm are described on Table 3.

Table 3: Algorithms Comparison

Algorithm	**Centroid Position**					**DB**	**Intra C.**	**Inter C.**	**Time**
		var1	**var2**	**var3**	**var4**	**Index**	**Dist.**	**Dist.**	**(s)**
Genetic	C1	12.647	28.967	2.562e-06	1.087	0.612	10.359	30.128	59.217
Algorithm	C2	4.461	3.531e-07	1.188	0.634		8.081		
Differential	C1	17.431	31	0.661	0	0.605	10.370	30.142	55.110
Evolution	C2	4.851	1.496	0.414	0		8.079		
PSO	C1	16.802	30.637	8.687e-08	0.374	0.598	10.989	31.382	42.821
	C2	3.866	2.0743	1.205	0.865		7.806		
k-means	C1	11.380	18.380	0.952	1.095	0.901	18.326	16.066	1.631
	C2	4.974	3.653	0.858	0.653		26.467		

Since in this paper, the automatic clustering algorithms are considered, it is up to the algorithm itself to define the optimal number of cluster division. In this case, the optimal value corresponds to the smallest DB-index, since the optimization algorithm goal is to minimize this parameter. Thence, from the results presented in Table 3, it is possible to observe that all algorithms pointed to 2 as the optimal cluster division number. That is, 2 cluster is the value that minimizes the DB-index for all considered bio-inspired clustering algorithms (GA, DE and PSO) and also by the Matlab function *evalclusters*, which was used to define the cluster number of the k- means.

From the results presented in Table 3 it can be said that the 3 evolutionary bio-inspired algorithms have similar behavior, both in the definition of the position of the centroids and also in the parameters value of DB index, intra-cluster distance and inter-cluster distance. However, the PSO bio-inspired clustering algorithm presented slightly better solutions, having obtained the lowest value of DB index and greater inter-clustering distance in less time than the GA and DE. PSO is one of the most famous bio-inspired algorithm due to its high exploration capacity, simplicity coding, and especially the high speed of convergence.

Such features were also evident from the results obtained in this work. Since a small size and low complexity dataset was considered, the similarity between the results of the bio-inspired algorithms is according to what was expected. As the complexity of the data increases, it is expected to find different amounts of clusters in each algorithm.

Regarding k-means, although it provides the solution in much less computational time than the other algorithms, the final solution is worst compared to the 3 bio-inspired algorithms in terms of DB index and also in relation to intra-cluster and inter-cluster distances. It is important to highlight that the DB index used in k-means was obtained by the Matlab function *evalclusters*, since k-means is not an example of automatic clustering, so it requires a specific technique to define the initial parameter k.

Due to the better performance of PSO algorithm its solution was chosen to be presented and analyzed. As it is no possible to represent a 4-dimensional graphic, Figure 3 presents the clustering division, according to 3 to 3 variable combination.

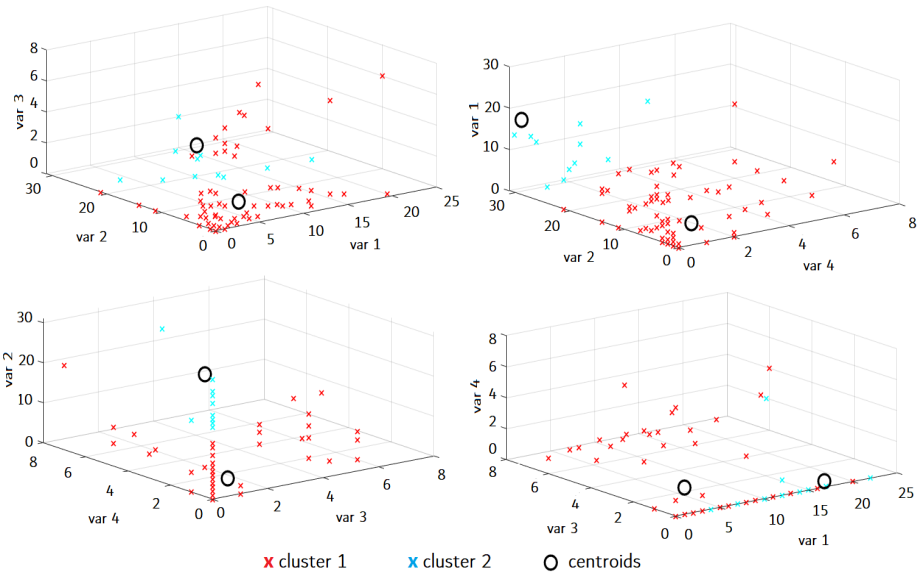

Fig. 3: PSO's bio-inspired clustering solution

Cluster 1, in blue, describes a group of students who answer more basic questions incorrectly. This cluster contains fewer elements, and it is slightly more compact than cluster 2. All students that belong to this cluster answered at least 18 basic questions incorrectly. However, it is not possible to establish an average value of basic questions answered by students in this cluster since the cluster elements are well scattered. Clearly, the characteristic of having at

least 18 incorrect answers is the main point in establishing the division between cluster 1 and 2. Thus, it can be said that most basic questions were answered incorrectly. Moreover, most of the students in this cluster do not answer any advanced questions, and of those who do, only take one test of 7 questions in the SNA. Out of the answers, either they get 2 correctly and 5 incorrectly, or they answer all incorrectly.

Cluster 2, in red, has the group of students who made fewer mistakes in basic questions, that is, less than 15, but it is essential to consider that they answered fewer questions than students from cluster 1. In general, students from cluster 2 answer around 20 basic questions; they usually take 2 tests of 7 questions in the SNA, with half of these answers being correct. On the other hand, concerning the advanced questions, they usually answer 1 or at most 2 tests in the SNA, and of the questions answered, they usually provide 5 correct answers.

6 Conclusions

E-learning has already operated a transformation in higher education, and online platforms such as MathE are an opportunity to make learning more accessible, deepen student engagement and allow teachers to shift to a student-centered pedagogical model. This work aimed to evaluate the performance of a group of students who answered questions about Linear Algebra on the section Self Need Assessment of the MathE platform. For this purpose, data collected over 3 years was evaluated through different cluster techniques.

Through the performed analysis it can be concluded that most of the students who use the MathE platform, specifically on the Linear Algebra topic, have many difficulties in the subject, as they have a high error rate about the hit rate. Although the clustering algorithm separates the sample into two groups, it was not possible to establish a group of students whose performance was significantly better than the other's. Besides, the expressive number of incorrect answers indicates that it is urgent and mandatory to review the questions' difficulty level. However, it is also known that some teachers use the platform in the classroom to ascertain the level of the students at the beginning of the course. This may be the cause of the high number of questions incorrectly answered since many of them are answered before the students have contact with the concepts in the classroom.

Future research will focus on developing a more robust clustering analysis and new possibilities in combining bio-inspired algorithms. Besides, more of the topics covered by the MathE platform must be involved in the study as well as other students' features such as country and course information.

References

1. Arthur, D., Vassilvitskii, S.: K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. p. 1027–1035. SODA '07, Society for Industrial and Applied Mathematics, USA (2007). https://doi.org/10.1145/1283383.1283494

2. Atabay, H.A., Sheikhzadeh, M.J., Torshizi, M.: A clustering algorithm based on integration of k-means and pso. In: 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC2016) - Higher Education Complex of Bam. pp. 59–63. Iran (2016). https://doi.org/10.1109/CSIEC.2016.7482110
3. Azevedo, B.F.: Study of Genetic Algorithms for Optimization Problems. Master's thesis, Instituto Politecnico de Braganca Escola Superior de Tecnologia e Gestao, Portugal, Braganca, Portugal (2020)
4. Azevedo, B.F., Amoura, Y., Kantayeva, G., Pacheco, M.F., Pereira, A.I., Fernandes, F.P.: Collaborative Learning Platform Using Learning Optimized Algorithms, vol. 1488. Springer (2021). https://doi.org/10.1007/978-3-030-91885-9-52
5. Azevedo, B.F., Pereira, A.I., Fernandes, F.P., Pacheco, M.F.: Mathematics learning and assessment using mathe platform: A case study. Education and Information Technologies . https://doi.org/10.1007/s10639-021-10669-y
6. Bansal, J.C., Singh, P.K., Nikhil, R.P.: Evolutionary and swarm intelligence algorithms. Studies in Computational Intelligence (2019). https://doi.org/10.1007/978-3-319-91341-4
7. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979). https://doi.org/10.1109/TPAMI.1979.4766909
8. Eesa, A.S., Orman, Z.: A new clustering method based on the bio-inspired cuttlefish optimization algorithm. Expert Systems **37** (2020). https://doi.org/10.1111/exsy.12478
9. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95 - International Conference on Neural Networks. vol. 4, pp. 1942–1948 vol.4 (1995). https://doi.org/10.1109/ICNN.1995.488968
10. Kuo, R.J., Amornnikun, P., Nguyen, T.P.Q.: Metaheuristic-based possibilistic multivariate fuzzy weighted c-means algorithms for market segmentation. Applied Soft Computing Journal **96** (2020). https://doi.org/10.1016/j.asoc.2020.106639
11. Kuo, R.J., Huang, Y.D., Lin, C.C., Wu, Y.H., Zulvia, F.E.: Automatic kernel clustering with bee colony optimization algorithm. Inf. Sci. **283**, 107–122 (2014). https://doi.org/10.1016/j.ins.2014.06.019
12. MATLAB: The mathworks inc. `https://www.mathworks.com/products/matlab.html` (2019a)
13. Nakane, T., Bold, N., Sun, H., Lu, X., Akashi, T., Zhang, C.: Application of evolutionary and swarm optimization in computer vision: a literature survey. IPSJ Transactions on Computer Vision and Applications **12**(3) (2020). https://doi.org/10.1186/s41074-020-00065-9
14. Nemmich, M.A., Debbat, F., Slimane, M.: A data clustering approach using bees algorithm with a memory scheme. Lecture Notes in Networks and Systems **50**, 261–270 (2019). https://doi.org/10.1007/978-3-319-98352-3-28
15. Nguyen, T.P.Q., Kuo, R.J.: Automatic fuzzy clustering using non-dominated sorting particle swarm optimization algorithm for categorical data. IEEE Access **7**, 99721–99734 (2019). https://doi.org/10.1109/ACCESS.2019.2927593
16. Pacifico, L.D.S., Ludermir, T.B.: An evaluation of k-means as a local search operator in hybrid memetic group search optimization for data clustering. NATURAL COMPUTING **20**(3, SI), 611–636 (2021). https://doi.org/10.1007/s11047-020-09809-z
17. Pedró, F., Subosa, M., Rivas, A., Valverde, P.: Artificial intelligence in education : challenges and opportunities for sustainable development (2019), uNESCO DOC Digital Library - Available online at https://unesdoc.unesco.org/ark:/48223/pf0000366994 accessed May, 2021

18. Qaddoura, R., Faris, H., Aljarah, I.: An efficient evolutionary algorithm with a nearest neighbor search technique for clustering analysis. Journal of Ambient Intelligence and Humanized Computing **12**, 8387–8412 (2021). https://doi.org/10.1007/s12652-020-02570-2
19. Saitta, S., Raphael, B., Smith, I.F.C.: A comprehensive validity index for clustering. Intell. Data Anal. **12**(6), 529–548 (2008). https://doi.org/10.3233/IDA-2008-12602
20. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory To Algorithms. Cambridge University Press (2014)
21. Singh, T.: A novel data clustering approach based on whale optimization algorithm. Expert Systems **38**(3) (2021). https://doi.org/10.1111/exsy.12657
22. Sivanandam, S.N., Deepa, S.N.: Introduction to Genetic Algorithms. Springer, 1 edn. (2008). https://doi.org/10.1007/978-3-540-73190-0
23. Storn, R., Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization **11**(4), 341–359 (1997). https://doi.org/doi.org/10.1023/A:1008202821328
24. Yapiz: Evolutionary clustering and automatic clustering. `https://www.mathworks.com/matlabcentral/fileexchange/52865-evolutionary-clustering-and-automatic-clustering` (2022), retrieved February 2, 2022.