Improving the Clustering Algorithms Automatic Generation Process with Cluster Quality Indexes

Michel Montenegro(✉), Aruanda Meiguins, Bianchi Meiguins, and Jefferson Morais

Computer Science Postgraduate Program, Federal University of Pará, Belém 66075-110, Brazil
michel.montenegro@gmail.com, aruandasimoes@gmail.com, {bianchi,jmorais}@ufpa.br

Abstract. AutoClustering is a computational tool for the automatic generation of clustering algorithms, which combines and evaluates the main parts of density-based algorithms to generate more appropriate solutions for a given dataset for clustering tasks. AutoClustering uses the Estimation of Distribution Algorithms (EDA) evolutionary technique to create the algorithms (individuals), and the adapted CLEST method (originally determines the best number of groups for a dataset) to compute individual fitness, using a decision-tree classifier. Thus, as the motivation to improve the quality of the results generated by AutoClustering, and to avoid possible bias by the adoption of a classifier, this work proposes to increase the efficiency of the evaluation process by the addition of a quality metric based on a fusion of three quality indexes of solution clusters. The three quality indexes are Silhouette, Dunn, and Davies-Bouldin, which assess the situation Intra and Inter clusters, with algorithms based on distance and independent of the generation of the groups. A final score for a specific solution (algorithm + parameters) is the average of normalized quality metric and normalized fitness. Besides, the results of the proposal presented solutions with higher cluster quality metrics, higher fitness average, and higher diversity of generated individuals (clustering algorithms) when compared with traditional Autoclustering.

Keywords: AutoClustering · Cluster quality index · Clustering algorithms

1 Introduction

Cluster analysis or clustering is a popular unsupervised learning technique used as an important task in the exploratory analysis of data when few or no prior knowledge is available [1]. Thus, the clustering task purpose is grouping data in

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12249, pp. 1017–1031, 2020.
https://doi.org/10.1007/978-3-030-58799-4_73

such a way that data from the same group (called a cluster) share similar characteristics relevant to the problem domain. At the same time, they are different or unrelated to data from other groups.

There is a wide variety of clustering algorithms described in the literature, each with its characteristics and peculiarities. The diversity of strategies for specifying clustering algorithms causes the development of extensions and variations of a set of fundamental algorithms, making the task of selecting the most appropriate algorithm even more complicated, which also considers specific characteristics of the problem domain and the dataset. Due to the presented scenario, it is not feasible to manually analyze all algorithm solutions, and many times the most appropriate algorithm for a given project is not applied.

There are some initiatives described in the literature that can assist in the selection process of the most appropriate clustering algorithm for a given dataset [2–5]. Frameworks like multiobjective clustering with automatic k-determination (MOCK) [2] and multi-objective clustering ensemble algorithm (MOCLE) [3] implement multi-objective evolutionary algorithms in their solutions. [4] proposed a hybrid algorithm, called Hybrid Selection Strategy (HSS), for selection of clustering partitions, combining multi-objective clustering and partition selection techniques.

Unlike the works mentioned above, in which the output of evolutionary algorithms is a set of clusters selected for a specific dataset, [5] developed a computational tool, called AutoClustering, for automatic generation of clustering algorithms, based on the evolutionary approach Estimation of Distribution Algorithm (EDA). The Autoclustering uses a modification of the CLEST method [6] as its fitness function, using a classifier based on a C4.5 decision tree implemented by the J48 algorithm [7], to objectively measure quality (fitness) of the clusters generated by the clustering algorithm for a given dataset.

One of the future works pointed out by [8] is the improvement in the confidence of the classifier results, avoiding possible bias in the results presented by AutoClustering. In this context, two solutions stand out: adoption of quality metrics of clusters formed by the candidate solutions, such as Silhouette, Dunn, and Davies-Bouldin, and use and fusion of other classifiers results, such as KNN and MLP [7,9]. However, a classifier committee is computationally costly, adding learning time and memory constraints to the problem [10].

Thus, this paper proposes to include in the evaluation process (CLEST method) of clustering algorithms generated by the Autoclustering tool a quality index considering the merger of results from the clustering quality metrics Dunn Index (DI) [11], Silhouette Index (SI) [12] and Davies-Bouldin Index (DBI) [13], to improve the confidence and avoid any possible bias caused by a classifier in the fitness calculation. The metrics evaluate the situation Intra and Inter Groups, with algorithms based on distance and independent of the generation of the groups.

For the evaluation of the proposed approach, comparative experiments were carried out between the previous AutoClustering evaluation model and the proposed one. Four datasets were used, three are public domain and belonging to the UCI data repository [14]: Glass Identification (9 attributes and 149

instances) [15], ClevelandHeart Diseases (14 attributes and 298 instances) [16], Bupa Liver-disorders (7 attributes and 345 instances) [17], and the fourth dataset is a synthetic base called Synthetic-1 (3 attributes and 589 instances) [18]. The experiments were carried out with the following setup: for each dataset, ten rounds, where each round had 500 generations, and 50 individuals per generation. The results showed that AutoClustering with the new evaluation model successfully identified among the candidate solutions those with the best cluster quality indexes and fitness across the generations, and improve the diversity of generated individuals in the and rounds, including new clustering algorithms.

The paper is organized as follows. Section 2 presents the theoretical foundations, including an introduction to AutoClustering, metrics for cluster evaluation, and the Normalization technique applied, Sect. 3 presents the proposed approach, Sect. 4 presents experiments and results, and Sect. 5 presents final considerations and future works.

2 Background

2.1 AutoClustering

AutoClustering [5] is a computational tool developed with a focus on the automatic generation of clustering algorithms, which combines and evaluates parts of density-based algorithms to generate the most suitable solutions for a given dataset for clustering tasks.

The AutoClustering is based on an evolutionary computing technique called the Estimation of Distribution Algorithms (EDA) [19] and uses a Directed Acyclic Graph (DAG) as an auxiliary data structure to generate the clustering algorithms. In the DAG (Fig. 1), each node represents the main procedures of a density-based clustering algorithm already in the literature, called building blocks, and the edges of this graph are the connections between these procedures, considering its input and output parameters.

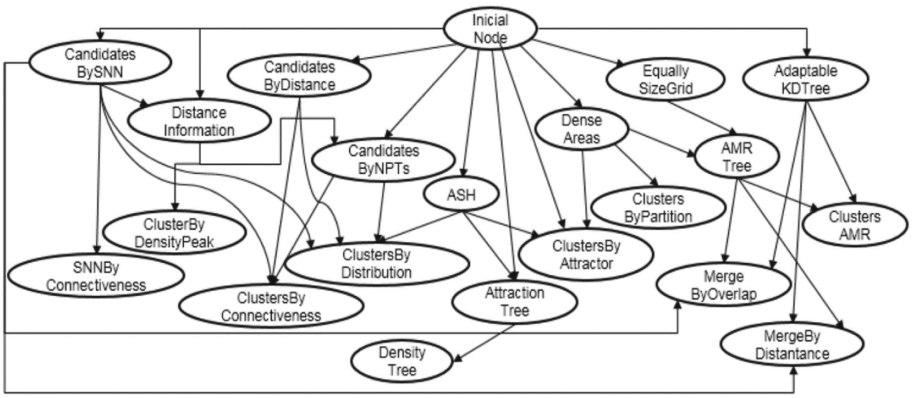

Fig. 1. Building blocks represented in DAG as auxiliary structure for EDA [5].

From the DAG, AutoClustering generates a population of individuals through the connection between these building blocks, with each block having its specific parameters. Each of the individuals (building blocks and parameters) represents a complete clustering algorithm. Figure 2 illustrates the individual generated from the DAG. The individuals generated can be clustering algorithms that already exist in the literature or new ones. The most recent version of AutoClustering includes 10 density-based clustering algorithms [8].

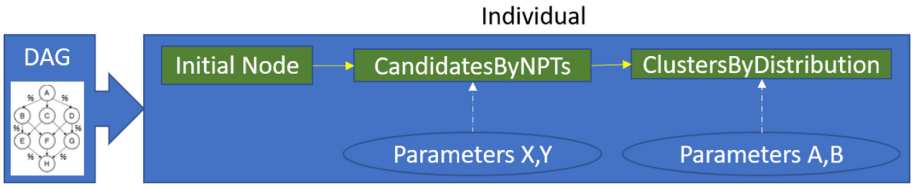

Fig. 2. Example of an individual referenced in a DAG. Adapted from [8].

The steps adopted in the execution of AutoClsutering to generate clustering algorithms are illustrated in Fig. 3.

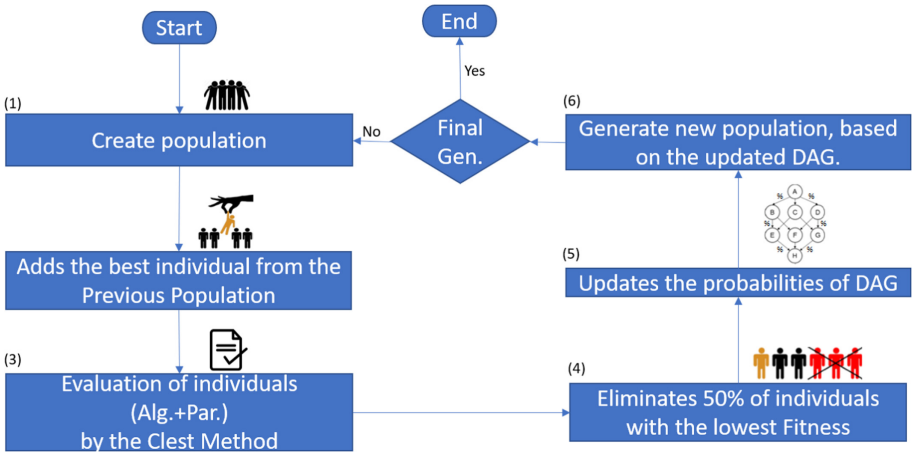

Fig. 3. Steps to run the AutoClustering tool. Adapted from [8].

AutoClustering uses a variation of the original CLEST method [6] to evaluate individuals. Its steps are:

1. Divide the dataset in training and testing.
2. Apply the clustering algorithm to the training base, to obtain a partition (designates labels for the base objects).

3. Apply a C classifier to the training base, using the previous labels as the class for each object.
4. Apply item 2 to the test base.
5. Apply the classifier C already trained in item 3, on the test base.
6. Compare the results of the classifier with the clustering algorithm.

The CLEST method uses a clustering algorithm (blocks + parameters) and a classifier (J48) to perform the assessment (cross-validation), the result of this assessment (Item 6) is the AutoClustering Fitness.

2.2 Cluster Quality Indexes

Silhouette Index: The silhouette index validates the clustering performance based on the difference in pairs between intra-cluster and inter-cluster distances. Moreover, the precise number is determined by maximizing the value of this index [22,28]. In the construction of the SI two items are necessary: the partition obtained by applying some clustering technique and the collection of all distances between objects. The Silhouette index is defined by:

$$S(A_i) = \frac{b(A_i) - a(A_i)}{\max(a(A_i), b(A_i))} \tag{1}$$

where, i represents each object into the same cluster A. For each object i it is calculated the value $S(A_i)$. $a(A_i)$ is the average distance between i and all other objects in the same cluster. $b(A_i)$ selects the lowest value of $d(A_i, C)$, and its definition is given by:

$$b(A_i) = \min(d(A_i, C)),\ para\ (C \neq A) \tag{2}$$

where, $d(A_i, C)$ is the average distance between i and all other objects on the other clusters C $(C \neq A)$.

The values of the silhouette index are in the range $-1 \leq S(A_i) \leq 1$, with the highest value being the optimal case for clustering [20].

Dunn Index: The dunn index identifies groups of clusters that are compact and well separated. This index is given by:

$$D(U) = \min\left(\frac{\min\left(\delta(A_i, A_j)\right)}{\max\{\Delta(C_k\})}\right) \tag{3}$$

where $i \ldots j$ is the interval between the objects into the cluster A and k is the objects in cluster as a whole C. $\delta(A_i, A_j)$ is the inter-cluster distance from A_i to A_j. $\Delta(C_k)$ is the distance intra-cluster of c_k. Finally, U is the partition where the clusters are located, for U↔A: $A_1 \cup A_i \cup A_c$, where A_i represents the cluster of that partition and c is the cluster number on the U partition.

The Dunn index aims to maximize the inter-cluster distance and minimize the intra-cluster distance. So the higher the value of D the better the result [20].

Davies-Bouldin Index: The Davies-Bouldin index identifies the sets of clusters that are compact and with good separation. This index is given by:

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c}\max_{i\neq j}\left(\frac{\Delta(A_i)+\Delta(A_j)}{\delta(A_i,A_j)}\right) \tag{4}$$

The lower the value of the Davies-Bouldin index (close to zero) more compact the clusters are and with more separate centers [20].

2.3 Normalization

Normalizing is a way of harmonizing values on different scales. A technique that maintains the relationship of normalized values to the original values is called Min-Max Normalization [31]. This technique adjusts values from predefined limits. Min-max normalization is given by:

$$S = \left(\frac{v - min(v)}{max(v) - min(v)}\right) * (max_s - min_s) + min_s \tag{5}$$

where, v is the current input value, $max(v)$ is the highest possible value for v and $min(v)$ is the lowest possible value for v. The minimum and maximum value desired as a return to the equation is represented respectively by min_s and max_s (for example, 0 and 1).

3 Proposed Approach

The proposed approach (Fig. 4) consists of combining of the CLEST method currently implemented in Autoclustering with Cluster assessment metrics Dunn, Silhouette, and Davies-Bouldin. More specifically, four more steps are added in the calculation of the individual's fitness of AutoClustering.

The first step implemented (item 4 of Fig. 4) performs the obtain of the clusters applying the generated individual on the dataset. After that (item 5), the clusters are evaluated using the metrics.

The third step added (item 6) consists of normalizing, using the Min-Max Method, the outputs from the CLEST method, and the evaluation metrics. This step is essential to standardize the values in the range of 0 to 1.

Finally, the fourth step of the proposal (item 7) generates the individual's final fitness value by calculating a simple average between the normalized values of the CLEST and the evaluation metrics.

It is important to note that any other metrics for assessing cluster quality can be added to the proposed approach. In other words, the main contribution of this work is to provide a baseline to be followed, keeping in mind that different techniques of normalization and cluster evaluation can be applied.

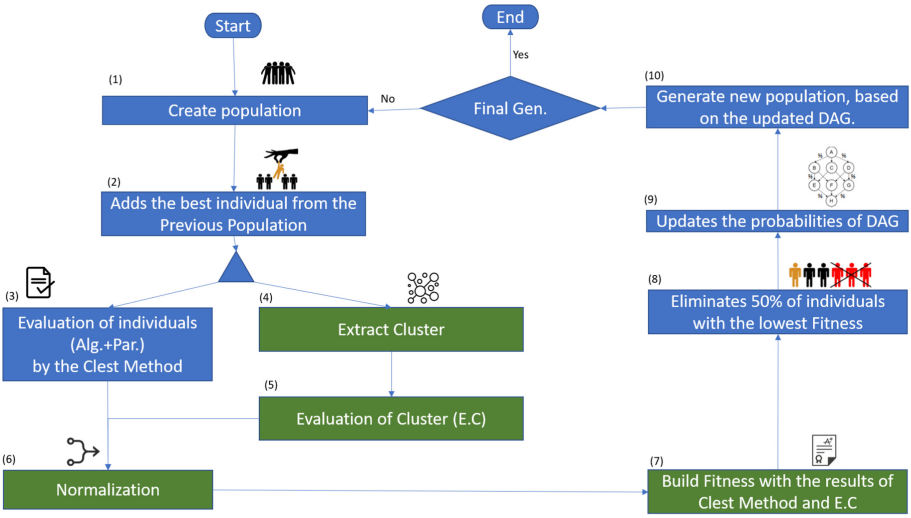

Fig. 4. General scheme of the proposed approach for calculating fitness in Autoclustering considering cluster assessment metrics.

4 Experiments and Results

4.1 Experimental Setup

Autoclustering was applied to four datasets, three datasets are public domain datasets and belong to the UCI data repository [14], they are: Glass Identification (9 attributes and 149 instances) [15], ClevelandHeart Diseases (14 attributes and 298 instances) [16], Bupa Liver-disorders (7 attributes and 345 instances) [17], and the fourth dataset is a synthetic base called Synthetic-1 (3 attributes and 589 instances) [18]. In general, the datasets are geared towards classification task, but the class attribute, when existing, was eliminated from the datasets before any Autoclustering processing. For each dataset, AutoClustering, with and without the proposed approach, was performed 10 times, with 500 generations per round, with a population of 50 individuals.

4.2 Results

In this section, some results will be presented considering the datasets used in the experiments showing the efficiency of the use of cluster quality indexes in the selection process and generation of clustering algorithm by Autoclustering.

Figure 5 and Fig. 6 show, respectively, alluvial graphics for traditional Auto-Clustering and the new evaluation model implemented in Autoclustering. It can be seen in the charts that the behavior of the probability models for the two proposals, considering the BUPA base and 500 generations of one round, both were updated differently, resulting in a greater number of different algorithms for the proposed model. For instance, the model proposed in Fig. 6 there are algorithms with three blocks, while in the traditional version, the best solutions have occurrences of only two blocks.

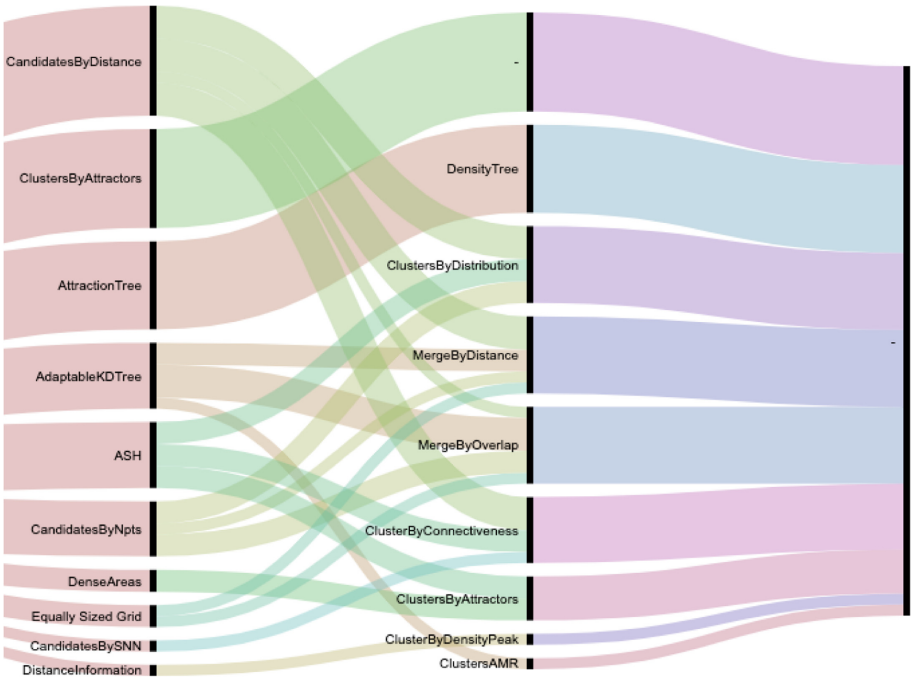

Fig. 5. Alluvial graph of the behavior of the Autoclustering probability model.

Figure 7 shows the fitness averages of individuals with non-zero fitness for each generation, considering all rounds to avoid possible bias. It is possible to observe that the average fitness for the proposed evaluation model was higher for all databases, which led to the choice of better individuals to update the AutoClustering probability model for the next generations. In other words, for the next generations, selected individuals were not only individuals with the best fitness but also with relevant group quality indexes.

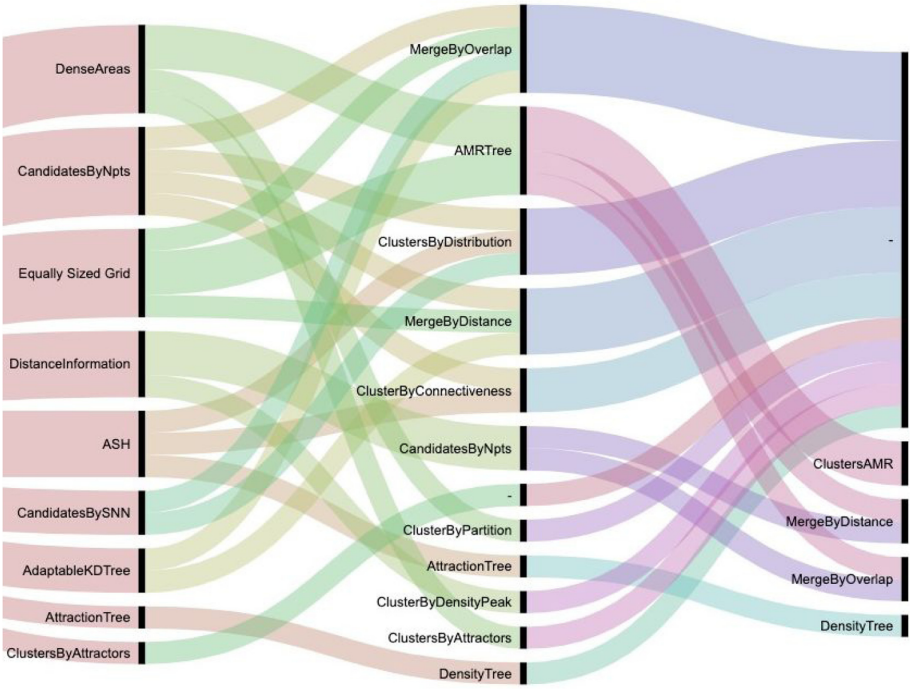

Fig. 6. Alluvial graph of the behavior of the probability model for the proposed approach.

Figure 8 shows a boxplot of Glass dataset considering all individuals from the last generation of each round (R1 ... R10) who had a fitness greater than zero. Analyzing Fig. 8, it can be seen that around 40% of the results the quality indexes contributed to a higher average of fitness. Also, in round 5 the proposed model presented the highest median and average of the results, in addition to other good results in rounds 9 and 10. This proves that the indexes adopted together with the evolutionary approach of the tool can converge to excellent results.

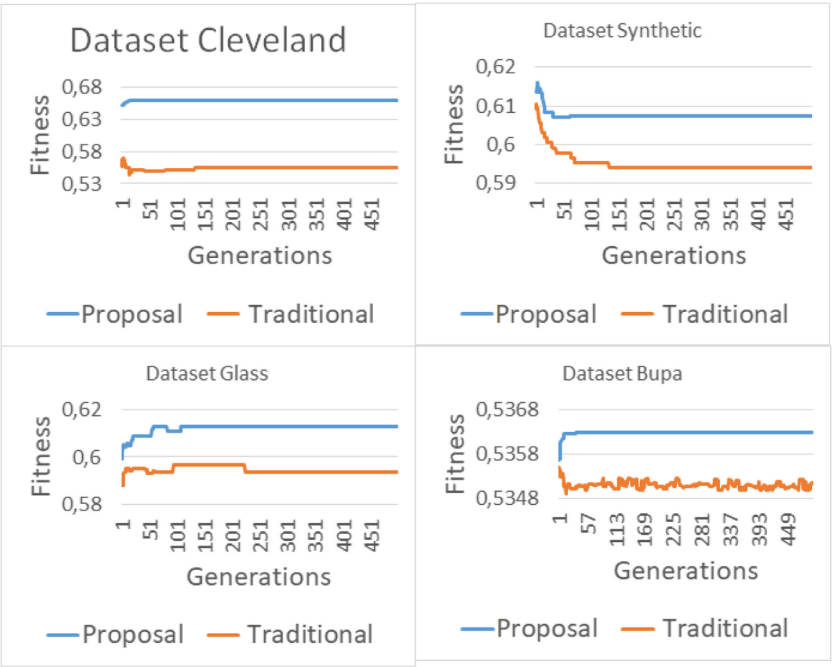

Fig. 7. Fitness averages of individuals with non-zero fitness for each generation.

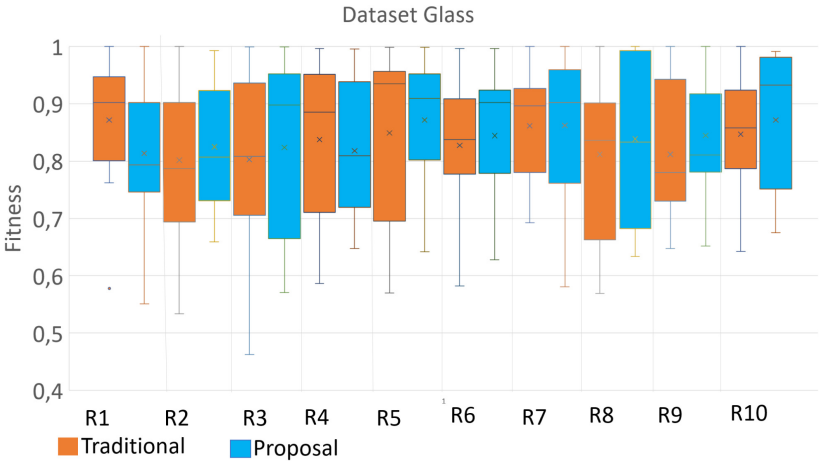

Fig. 8. Fitness boxplot of the proposed approach and traditional Autoclustering considering individuals with fitness greater than zero from the last generation of each round.

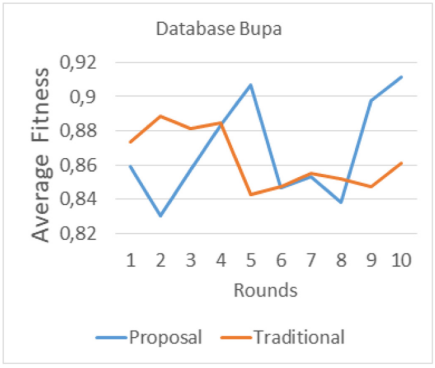

Fig. 9. Bupa - average per rounds.

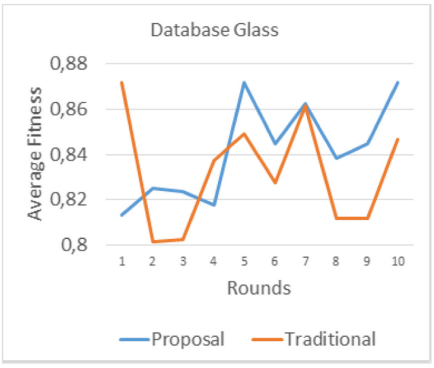

Fig. 10. Glass - average per rounds.

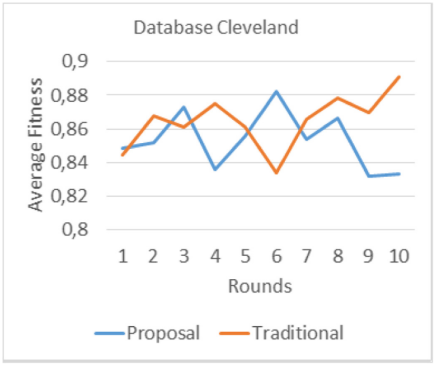

Fig. 11. Cleveland - average per rounds.

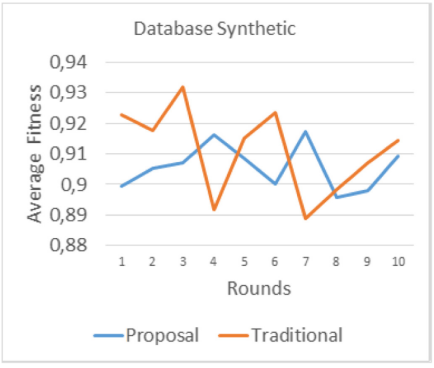

Fig. 12. Synthetic - average per rounds.

Another important point to highlight is that other individuals with fitness different of zero value were considered in the evolutionary process, which can lower the average of the values, however increasing the diversity of individuals throughout the evolutionary process, already presented in Fig. 5 and Fig. 6, and individuals who had high fitness values were preserved. Thus, the quality indexes applied do not change the objective of AutoClustering to generate more suitable individuals to the dataset, but contributed to an increase in the variability and quality of individuals in each generation, and therefore in the final result (Fig. 9, 10, 11 and 12).

For the visualization and comparison of the individuals generated at the end of a round, considering 500 generations for the Bupa dataset, the treemap technique presented in Fig. 13 is used. This technique allows to group data by attributes of the dataset, and at first, the results of both proposals are grouped by attribute named "tool" (traditional represented by the orange area and proposal represented by the blue area). The color of the blocks represents the fitness values, where the darkest represents the highest fitness values, and the size of

the rectangles represents the occurrence of the algorithm types in the last generation. Besides, there is a classification if the generated algorithm is a traditional algorithm or a new one, depending on blocks connections. It is possible to perceive again a higher number of generated algorithms, a greater diversity, a higher quantity of new algorithms created, as well as better individuals in the results for the AutoClustering with the new evaluation model. Thus, the application of quality indexes improved the diversity in the results of Autoclustering without losing reference to the best fitness values.



Fig. 13. Treemap for visualization and comparison of individuals generated at the end of a round, considering 500 generations for the Bupa dataset. (Color figure online)

5 Final Remarks and Future Works

This work aimed to propose a new model for evaluating the selection of individuals in the evolutionary process of the generations of the AutoClustering tool. This selection process is calculated by the adapted CLEST method, and grouping quality indexes have been added to compose a new quality metric, which is the average fitness calculated by CLEST and the average of three group quality indices (Silhouette, Dunn, and Davies-Bouldin). The results showed that the approach is promising because it allows an increase in the diversity and quality of individuals in each generation of the evolutionary algorithm used as a usage scenario.

Nevertheless, further investigation is needed. Future works include to evaluate other clustering quality indexes; to evaluate other classifiers together with, or replace, the CLEST method; to investigate the impacts when considering the only one of the cluster quality indexes or a combination of them; finally to evaluate other scenarios using different evolutionary tools and datasets with similar proposals.

Acknowledgments. This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior - Brasil (CAPES), under the Program PROCAD-AMAZÔNIA, process nº 88881.357580/2019-01.

References

1. Han, J., Kamber, M., Pei, J.: Data Mining Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers (2012). http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf. Accessed 17 Apr 2020
2. Matake, N., Hiroyasu, T., Miki, M., Senda, T.: Multiobjective clustering with automatic k-determination for large-scale data. In: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, pp. 861–868 (2007) https://doi.org/10.1145/1276958.1277126
3. Faceli, K., Souto, M.C.P., Araújo, D.S.A., Carvalho, A.C.P.L.F.: Multi-objective clustering ensemble for gene expression data analysis. J. Neurocomput. **72**(13–15), 2763–2774 (2009). https://doi.org/10.1016/j.neucom.2008.09.025
4. Antunes, V., Sakata, T.C., Faceli, K., Souto, M.: Hybrid strategy for selecting compact set of clustering partitions. Appl. Soft Comput. **87**, 105971 (2020). https://doi.org/10.1016/j.asoc.2019.105971
5. Meiguins, A.S.G., Limao, R.C., Meiguins, B.S., Junior, S.F.S., Freitas, A.A.: Auto-Clustering: an estimation of distribution algorithm for the automatic generation of clustering algorithms. In: Proceedings of WCCI 2012 - IEEE World Congress on Computational Intelligence (Congress on Evolutionary Computation), pp. 2560–2566. IEEE Press (2012)
6. Dudoit, S., Fridlyand, J.: A prediction-based resampling method for estimating the number of clusters in a data set. Genome Biol. **3**(7), research0036.1 (2002) https://doi.org/10.1186/gb-2002-3-7-research0036
7. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2016)
8. Meiguins, A., Santos, Y., Santos, D., Meiguins, B., Morais, J.: Visual analysis scenarios for understanding evolutionary computational techniques' behavior. Inf. Open Access J. **10**(3), 88 (2019). https://doi.org/10.3390/info10030088
9. Wu, X., et al.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2007). https://doi.org/10.1007/s10115-007-0114-2
10. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. **33**, 1–39 (2010). https://doi.org/10.1007/s10462-009-9124-7
11. Rendon, E., Abundez, I., Arizmendi, A., Quiroz, E.M.: Internal versus external cluster validation indexes. Int. J. Comput. Commun. **5**(1), 27–34 (2011). https://www.researchgate.net/profile/Erendira_Rendon. Accessed 17 Apr 2020
12. Rendon, E., Abundez, I., Gutierrez, C., Dáaz, S.: A comparison of internal and external cluster validation indexes. In: Proceedings of the 2011 American Conference, San Francisco, CA, USA (2011). https://www.researchgate.net/profile/Erendira_Rendon. Accessed 17 Apr 2020
13. Sanchez, M.S., Valdovinos, R.M., Trueba, A., Rendon, E., Alejo, R., Lopez, E.: Applicability of cluster validation indexes for large data sets. In: 12th Mexican International Conference on Artificial Intelligence (2013). https://doi.org/10.1109/micai.2013.30

14. Lichman, M.: UCI Machine Learning Repository. University of California, Oakland, CA, USA (2013)
15. Spiehler, V.: From USA Forensic Science Service; 6 types of glass; defined in terms of their oxide content (i.e. Na, Fe, K, etc). http://archive.ics.uci.edu/ml/datasets/glass+identification. Accessed 25 Apr 2020
16. Aha, D.W.: 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach. http://archive.ics.uci.edu/ml/datasets/Heart+Disease. Accessed 25 Apr 2020
17. Forsyth R.S.: BUPA Medical Research Ltd. http://archive.ics.uci.edu/ml/datasets/liver+disorders. Accessed 25 Apr 2020
18. Brito, Y.P.S., Santos, C.G.R., Mendonça, S.P., Aráujo, T.D., Freitas, A.A., Meiguins, B.S.: A prototype application to generate synthetic datasets for information visualization evaluations. In: 2018 22nd International Conference Information Visualisation (IV), pp. 153–158 (2018)
19. Ceberio, J., Irurozki, E., Mendiburu, A., et al.: A review on estimation of distribution algorithms in permutation-based combinatorial optimization problems. Prog. Artif. Intell. **1**, 103–117 (2012). https://doi.org/10.1007/s13748-011-0005-3
20. Bolshakova, N., Azuaje, F.: Cluster validation techniques for genome expression data. Sig. Process. **83**(4), 825–833 (2003). https://doi.org/10.1016/s0165-1684(02)00475-9
21. Kim, T.K.: T test as a parametric statistic. Korean J. Anesthesiol. **68**(6), 540 (2015). https://doi.org/10.4097/kjae.2015.68.6.540
22. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: IEEE International Conference on Data Mining (2010). https://doi.org/10.1109/icdm.2010.35
23. Tan, P.N., Kumar, V., Steinbach, M.: Introduction to Data Mining, 1st edn. Pearson Addison Wesley (2006). http://repository.fue.edu.eg/xmlui/bitstream/handle/123456789/3583/8857.pdf. Accessed 12 Apr 2020
24. Meiguins, A.S.G., Freitas, A.A., Limão, R.C., Junior, S.F.S., Meiguins, B.S.: An estimation of distribution algorithm for the automatic generation of clustering algorithms. In: Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation - GECCO 2010 (2010). https://doi.org/10.1145/1830483.1830679
25. Frank, A., Asuncion, A.: UCI Machine Learning Repository. Journal Irvine, University of California, School of Information and Computer Science, CA (2010). http://archive.ics.uci.edu/ml. Accessed 05 Apr 2020
26. Cole, T.J.: Too many digits: the presentation of numerical data: Table 1. Arch. Dis. Child. **100**(7), 608–609 (2015). https://doi.org/10.1136/archdischild-2014-307149
27. Tiwari, R., Singh, M.: Correlation-based attribute selection using genetic algorithm. Int. J. Comput. Appl. **4**(8) (2010). https://doi.org/10.5120/847-1182
28. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7
29. Ahuja, K., Saini, A.: Analyzing formation of K mean clusters using similarity and dissimilarity measures. Int. J. Adv. Trends Comput. Sci. Eng. - IJATCSE **2**(1), 72–74 (2013)
30. Bhm, R., Rockenbach, B.: The inter-group comparison - intra-group cooperation hypothesis: comparisons between groups increase efficiency in public goods provision. PLoS ONE **8**(2), e56152 (2013). https://doi.org/10.1371/journal.pone.0056152

31. Patro, S.G., Sahu, K.K.: Normalization: a preprocessing stage. Int. Adv. Res. J. Sci. Eng. Technol. - IARJSET **2**(3) (2015). https://doi.org/10.17148/IARJSET.2015.2305
32. Sedgwick, P.: Pearson's correlation coefficient. BMJ **345**, e4483–e4483 (2012). https://doi.org/10.1136/bmj.e4483