U-control Chart based Differential Evolution Clustering for Determining the Number of Cluster in k-Means

Jesús Silva[1], Omar Bonerge Pineda Lezama[2], Noel Varela[3], Jesús García Guiliany[4], Ernesto Steffens Sanabria[5], Madelin Sánchez Otero[6], Vladimir Álvarez Rojas[7]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
jesussilvaUPC@gmail.com
[2]Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras
omarpineda@unitec.edu
[3]Universidad de la Costa, Barranquilla, Colombia
nvarela2@cuc.edu.co
[4]Universidad Simón Bolívar, Barranquilla, Colombia
jesus.garcia@unisimonbolivar.edu.co
[5]Corporación Universitaria Latinoamericana, barranquilla Colombia
steffensse@ul.edu.co
[6]Corporación Universitaria Rafael Núñez, Cartagena Colombia
madelin.sanchez@curnvirtual.edu.co
[7]Corporación Universitaria Minuto de Dios- UNIMINUTO, Bello, Antioquia
vladimir.alvarez@uniminuto.edu

Abstract. The automatic clustering differential evolution (ACDE) is one of the clustering methods that are able to determine the cluster number automatically. However, ACDE still makes use of the manual strategy to determine k activation threshold thereby affecting its performance. In this study, the ACDE problem will be ameliorated using the u-control chart (UCC) then the cluster number generated from ACDE will be fed to k-means. The performance of the proposed method was tested using six public datasets from the UCI repository about academic efficiency (AE) and evaluated with Davies Bouldin Index (DBI) and Cosine Similarity (CS) measure. The results show that the proposed method yields excellent performance compared to prior researches.

Keywords: k-means, automatic clustering, differential evolution, k activation threshold, u control chart, academic efficiency (AE).

1. Introduction

The aim of this study is to apply the u-chart control (UCC) method to determine the k-activation threshold on Automatic Clustering Differential Evolution (ACDE) in order to identify behavior patterns and relations between the different attributes, enabling to identify and predict the likelihood of student desertion by foreseeing the factors influencing their permanence, generating knowledge for timely decisions, and offering competitive advantage to the institution where ACDE is used to determine the number of clusters in k-means automatically and improving the performance of k-means.

The k-means method is one of the hard partition methods in cluster analysis of the data mining field. The k-means has advantages, i.e. it is easy to implement grouping a large dataset, and with stable performance over different problems (Ben Salem et al., 2018 [1]; Chakraborty and Das, 2018 [2]). However, the clustering results of k-means depends on a certain number of clusters as inputs. If the estimated number of clusters does not tally with the final solution, the chances of clustering are very low (Abdul Masud et al., 2018 [3]; Rahman et al., 2015 [4]; Rahman and Islam, 2014 [5]; Ramadas et al., 2016 [6]). Meanwhile, getting the number of k as an input on k-means is still not an easy task because the user requires a prior specification number of the cluster (Yaqian et al., 2017 [7]). This condition is termed a local optimum problem (Tîrnăucă et al., 2018 [8]). In practice, the local optimum problem is overcome by applying the method several times with a different number of k, then choosing the best results. Determining the number of clusters is significant for the k-means method (Xiang et al., 2015 [9]). Automatic clustering methods are a solution that helps the user determine the optimal number of clusters (Garcia and Flores, 2016 [10]). Therefore, the automatic clustering method is an effective solution to this problem.

Researches on the determination of the number of clusters used automatic clustering methods which are based on the Evolutionary Computation (EC) technique. K-means method has done a lot and has been published with different methods, namely Automatic Clustering using Differential Evolution (ACDE) (Das et al., 2008 [11]), combining methods between PSO and k-means on Dynamic Clustering with Particle Swarm Optimization (DCPSO) (Omran et al., 2006 [12]), and Genetic Clustering for unknown k clustering (GCUK) (Bandyopadhyay and Maulik, 2002 [13]).

Automatic clustering methods have been used to determine the number of clusters in the k- means but are yet to achieve an accurate cluster result. Therefore, it is necessary to improve the performance of automated grouping methods used for determining the number of clusters. The ACDE method is the most popular EC technique which has effectively improved the performance of automatic clustering methods proposed by previous researchers (Das et al., 2008 [11]). ACDE predicated on the differential evolution (DE) method is one of the strongest, fastest, and most efficient global search heuristics methods in the world that is very easy to use with high-dimensional data. It can be employed using polynomial functions and other functions because it is easy to change the values of control variables such as NP, F, and CR to obtain good search results (Ramadas et al., 2016 [6]). However, ACDE has a weakness in determining the k activation threshold that is still dependent on user judgment (Tam et al., 2017 [14]).

The ACDE was then developed by (Kuo et al., 2013 [15]) and the combination of ACDE and k-means methods was termed the automatic clustering approach based on the differential evolution method combined with k-means for crisp clustering method aimed at improving clustering performance in the k-means method (ACDE-k-means). The ACDE method can find the number of clusters automatically and is able to balance the evolutionary process of DE methods to achieve better partitions than the classic DE. However, the classic DE method still depends on user's considerations to determine the k activation threshold thereby affecting the performance of the DE method (Piotrowski, 2017 [16]).

The U-Control Chart (UCC) method is employed to determine the k activation threshold that is used for the initial step to get the value of the variables sought before initialization of the variable vector. The UUC is a method from statistical process control (SPC) which has proved to be effective in solving the problem of management control attributes (Kaya, 2009 [17]). Other methods such us P-Control Chart and C-Control Chart are valid methods, but not used. This research focuses just on UCC. The UCC method used to average the data to be measured is then reduced and added to find upper and lower bound values on the number of attributes for the searched variables. A product is said to have a good quality if the average value is at a threshold or the average value is between the upper and lower bound. Based on the above assumption, the data is good if it is within the threshold of the U-Control Chart.

2. Theoretical Review

Several studies have been carried out to find the number of clusters of k-means on automatic clustering evolutionary methods. The most used clustering methods, that is, the main methods, combined with evolutionary computation methods, are DCPSO (Omran et al., 2006 [12]), GCUK (Bandyopadhyay and Maulik, 2002 [13]), ACDE-k-means (Kuo et al., 2013 [15]) and are often used for improving automatic clustering k-means results. So far, the clustering performance method to achieve optimal cluster number results is still a subject of further research because the best performance from all evaluations has not been completely achieved.

In [10], they proposed the Genetic Clustering for Unknown k (GCUK) method to find the number of automatic clusters k-means. This method begins by encoding several clusters premised on the prototype don't-care. Chromosome compatibility is determined by objective function using the DBI to guide evolutionary search (Bandyopadhyay and Maulik, 2002 [13]). The value of the valid partition result is the smallest value of the objective function (Das et al., 2008 [11]). It employs public datasets from the UCI machine learning repositories such as cancer, synthetic, spot image, and iris data. The result indicated that this method can automatically find few land-cover varieties even when the size of the data set is considerably large.

(Omran et al., 2006 [12]) proposed a combination of PSO and k-means for finding the number of clusters in an unlabeled data named Dynamic Clustering PSO (DCPSO). PSO is implemented to find the number of clusters in a data and k-means is implemented to do repair grouping. However, they are still limited to testing the original image data and some of the synthetic image data (Dobbie et al., 2014 [18]). The authors employ two different dataset types: natural images and synthetic image data (10 datasets) (Lenna, Mandrill, Jet, Peppers, MRI, and Tahoe) from UCI machine learning repository where each DBI average of sample natural images and this method can achieve a decrease in the performance of k-means.

A new method called ACDE-k-means proposed by Kuo et al. (2013) [15] is developed from a combination of ACDE and k-means method for crisp clustering. In this case, ACDE uses the basic DE method that has weaknesses as described earlier and differential evolution clustering is considered as automatic. The aim of this method is to find the optimal number of clusters in k-means without knowing the information from the data a priori. Also, the differential evolution clustering automatic efficiency for the high-dimensional dataset outperforms other related studies such as GCUK and DCPSO (Garcia and Flores, 2016) [10]. The two indexes of evaluation used are DBI and CS, then, dataset tested is from UCI repository: Breast Cancer, Iris, Vowel end

Glass, and Wine. Finally, the decreasing average value of DBI was determined and CS measure of each dataset of k-means.

3. Materials and methods

3.1 Database

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets. As an indication of the impact of the archive, it has been cited over 1000 times, making it one of the top 100 most cited "papers" in all of computer science. The current version of the web site was designed in 2007 by Arthur Asuncion and David Newman, and this project is in collaboration with Rexa.info at the University of Massachusetts Amherst. Funding support from the National Science Foundation is gratefully acknowledged [19]. The proposed method was tested using two different datasets, namely artificial and real datasets. Each dataset can be seen in Table 1.

Table 1. Description of the UCI database tables

Databases	Type of Dataset	Number of records	No. Attributes	Class (k)
Synthetic	Artificial	458	69	4
Wine Quality	Artificial	215	10	4
Abalone	Artificial	750	4	7
Poker Hand	Real	278	4	4
Glass	Real	1520	8	20
Iris	Real	576	27	5

3.2 Methods

The Objective function is a simulated search on a dataset to guide towards an optimal global solution. In the case of the clustering problems, the objective function usually uses the cluster validity index (Garcia and Flores, 2016) [10]. In this case, Davies Building index (DBI) and cosine similarity measure (CS) are used as objective function based on the finding of Das et al. (2008) [11] as follows Eq. (1) and Eq. (2).

$$f1 = \frac{1}{CS_i(K) + eps} \quad \textbf{(1)}$$

The eps is a small bias term equal to $2*10^{-6}$ near zero. $2*10^{-6}$ is a cluster k for k with set number of clusters as initialization to cluster of the datasets.

$$f2 = \frac{1}{DBI_i(K)+eps} \tag{2}$$

Where DBIi is the DB index, evaluated on the partitions yielded by the i-th vector and eps is the same as before.

In this research, a combination of the U-Control Chart (UCC) and Automatic Clustering using Differential Evolution method is proposed to determine the number of clusters on k-means. The aim of the UCC method is to control k activation threshold of the Automatic Clustering using the Differential Evolution method. The latter will automatically search the optimal number of clusters in the data as required by k-means. The representation of chromosome used is based on [12]. Because the Automatic Clustering using Differential Evolution method produces a premature cluster, the k-means is implemented to repair the premature clustering. As shown in Figure 1, the steps for the complete proposed method are given here.

Step 1. Prepare datasets.

Step 2. Initialize each chromosome containing a selected number of k randomly selected clusters and specify the k activation threshold using the UCC method defined with a stage as follows Eq. (3), (4) and (5). In Eq. (3), the average value is given by the average of all attributes. After that Eq. (4), the upper bound (ub) is calculated. Next Eq. (5), the lower bound (lb) is calculated.

$$\bar{u} = \sum x_i \tag{3}$$

$$ub = \bar{u} + K\sqrt{\frac{\bar{u}}{n_i}} \tag{4}$$

$$lb = \bar{u} - K\sqrt{\frac{\bar{u}}{n_i}} \tag{5}$$

Step 3. Generate initial population randomly based on predetermined k activation threshold values.

Step 4. Find the active cluster center, which is defined as shown in the following.

IF $V_{i,k}T_k > 0.5$ THEN cluster center $V_{i,k}m_k$ is ACTIVE
ELSE $V_{i,k}m_k$ is INACTIVE.

Where the center of the $V_{i,k}$ cluster on the chromosome will be active or selected if $V_{i,k}T_k > 0.5$. Conversely, if $V_{i,k}T_k < 0.5$ the center of the cluster $V_{i,k}$ is not active in the i-th chromosome. $V_{i,k}T_k$ is the cost population of the data generation, while the best solution, cost or $V_{i,k}m_k$ is the best solution for each iteration.

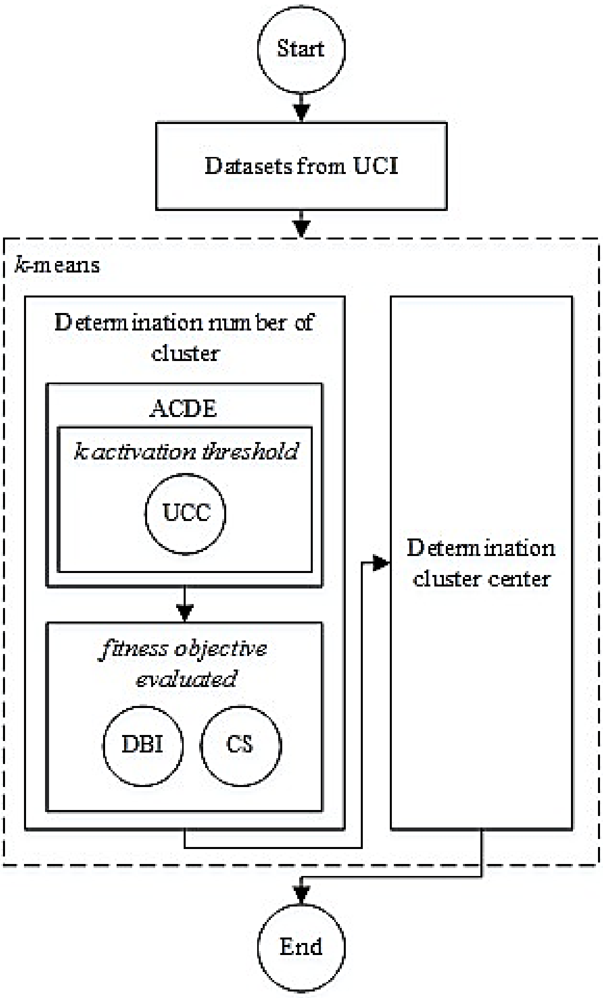

Start
Datasets from UCI
k-means
Determination number of cluster
ACDE
k activation threshold
UCC
fitness objective evaluated
DBI
CS
Determination cluster center
End

Fig 1. Block Diagram of the Proposed Method [16].

Step 5. For iteration is the best solution for each iteration

- Find the distance of each data vector from all active centroids of the i-th chromosome,
- Allocate each data vector to a cluster with the shortest distance,
- Change member(s) of the population (based on DE method) using the objective function to make the selected population better.
- Apply k-means method. The active cluster number is used as input k-means to adjust i-th active chromosome.

Step 6. The minimum objective is the output of the global best chromosome.

4. Results

The experiments were conducted using a computing platform with Intel Celeron 2.16 GHz CPU, 8 GB RAM and Microsoft Windows 10 Home 64-bit used as the operating system, and MATLAB version R2016a used as the data analytics tool. MATLAB would produce a model performance as the calculation output, such as the average value best cluster DBI and CS measure.

4.1 Data Transformation Stage

In the transformation stage, useful features are observed to represent the data depending on the goal of the data mining process. Methods are used for reducing dimensions or reducing the effective number of variables under consideration or to find invariant representations of the data [3].

At this stage, the UCI dataset was built, integrating the attributes of the different tables of the database (see Table 1). Items with missing data were suppressed, new attributes were built (see Table 2) and the continuous attributes were discretized the numerical values were transformed into discrete or nominal attributes. Some of the discretized attributes are shown in Tables 3, 4 and 5.

On the other hand, the UCI dataset was adapted to the ARFF format (Attribute, Relation, File, Format).

The structure of the ARFF format [13] is the following:

- Header: defines the name of the relationship and its format is as follows:
- @Relation <name-of-the-relationship>
- Statements by the attributes. In this section, the attributes that will compose the ARFF file with its type are declared. The syntax is as follows:
- @Attribute <attribute-name> <type>
- Data section. The data that make up the relationship between commas separating the attributes and with lines break in relationships are declared.

Table 2. Resume of new attributes in the data set UCI

Databases	Type of Dataset	Number of records	No. Attributes	Class (k)
Synthetic	Artificial	351	48	4
Wine Quality	Artificial	178	10	4
Abalone	Artificial	651	3	7
Poker Hand	Real	206	3	4
Glass	Real	1315	5	20
Iris	Real	508	19	5

Finally, the data set with 26 UCI.ARFF was obtained with 26 attributes and 20.329 records, ready to apply the data mining techniques, using the described methodology to obtain the patterns of low academic performance and/or desertion of students from the Colombian universities.

Table 3. Discretization of the attributes of Wine Quality

Wine Quality	**Value**	**No. Records**
Color	A	15
Aroma	B	20
Flavor	C	18
Ripening period	D	62

Table 4. Discretization of the attributes of Glass

Glass	**Value**	**No. Records**
Color	A	21
Air	B	40
Sulfur quality	C	30
Silicon quality	D	17

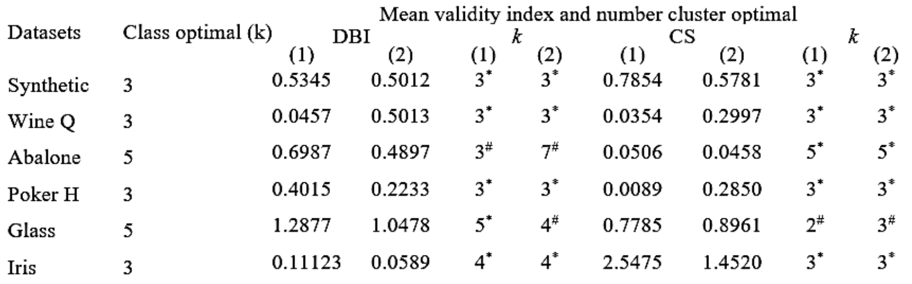
Table 5. Discretization of the attributes of Iris

Iris	**Value**	**No.Records**
Color	A	42
Age	B	10
Sex	C	12
Class	D	80

The proposed method was applied to the six databases listed in Table 1. The Parameter setting for proposed method based on the recommendation of Das et al. (2008) [13] is as follows: maxiter = 300, pop-size=10*dim, CRmax = 1.0 and CRmin = 0.45. Max-iteration indicates the amount of iteration, pop-size is the size of the population, cross-over probability is used to initialize the position of a particle or chromosome.

The best model automatic clustering on each dataset is highlighted with boldfaced print and the best optimal cluster result is marked with (1) and (2) squared on each dataset. As shown in Table 6, the second experiment UCC+ACDE-k-means is outperforming in almost all datasets with respect to DBI (5 of 6 datasets) and the optimal search k results of both methods are extremely good except for the vowel dataset. Results comparison ACDE-k-means only vs UCC+ACDE-k-means.

Table 6. Results comparison ACDE-k-means only vs UCC+ACDE-k-means.



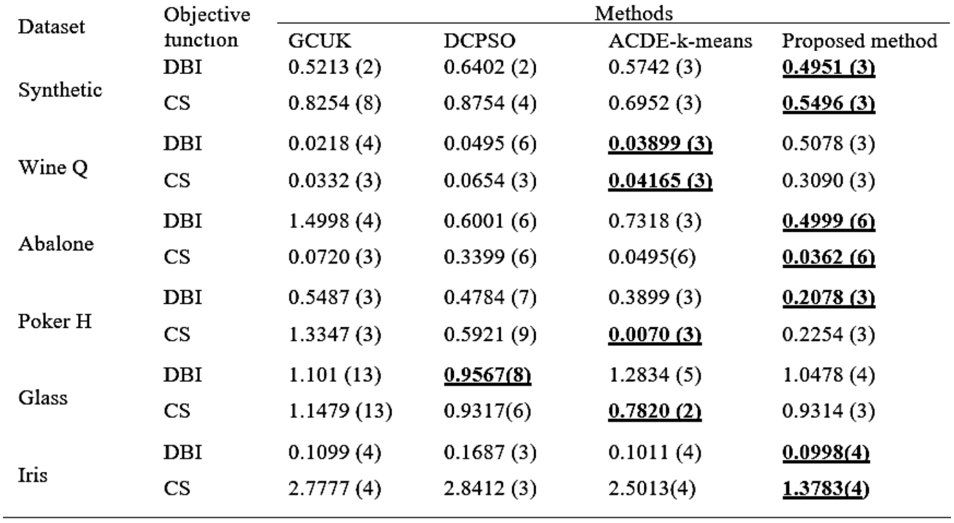
Datasets	Class optimal (k)	Mean validity index and number cluster optimal							
		DBI		*k*		CS		*k*	
		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Synthetic	3	0.5345	0.5012	3^*	3^*	0.7854	0.5781	3^*	3^*
Wine Q	3	0.0457	0.5013	3^*	3^*	0.0354	0.2997	3^*	3^*
Abalone	5	0.6987	0.4897	$3^\#$	$7^\#$	0.0506	0.0458	5^*	5^*
Poker H	3	0.4015	0.2233	3^*	3^*	0.0089	0.2850	3^*	3^*
Glass	5	1.2877	1.0478	5^*	$4^\#$	0.7785	0.8961	$2^\#$	$3^\#$
Iris	3	0.11123	0.0589	4^*	4^*	2.5475	1.4520	3^*	3^*

$^{(1)}$ACDE-k-means only ; $^{(2)}$UCC-ACDE-k-means ; *Number cluster optimal; $^\#$not optimal

Meanwhile, the first experiment (ACDE-k-means only) only outperforms in Wine Quality. Concerning CS, the first and the second experiment on each dataset produced excellent three different datasets. Meanwhile, regarding the determination of the number of k optimal, both methods are extremely good except in the Glass. However, based on this study, the overall second experiment outperformed and is better than the first experiment since the main evaluation method used in finding the number of the optimal cluster such as all six datasets used is DBI.

Finally, the proposed method was compared with prior researches such as Omran et al. (2006) [12] method, MAcQueen (1967) method, Kuo et al. (2013) [15] method as well as Bandyopadhyay and Maulik (2002) [13] method. Table 7 shows the comparison of prior research and proposed method with all datasets.

Table 7. Comparison to prior search based on DBI and CS as objective function for all datasets.



Dataset	Objective function	Methods			
		GCUK	DCPSO	ACDE-k-means	Proposed method
Synthetic	DBI	0.5213 (2)	0.6402 (2)	0.5742 (3)	**0.4951 (3)**
	CS	0.8254 (8)	0.8754 (4)	0.6952 (3)	**0.5496 (3)**
Wine Q	DBI	0.0218 (4)	0.0495 (6)	**0.03899 (3)**	0.5078 (3)
	CS	0.0332 (3)	0.0654 (3)	**0.04165 (3)**	0.3090 (3)
Abalone	DBI	1.4998 (4)	0.6001 (6)	0.7318 (3)	**0.4999 (6)**
	CS	0.0720 (3)	0.3399 (6)	0.0495(6)	**0.0362 (6)**
Poker H	DBI	0.5487 (3)	0.4784 (7)	0.3899 (3)	**0.2078 (3)**
	CS	1.3347 (3)	0.5921 (9)	**0.0070 (3)**	0.2254 (3)
Glass	DBI	1.101 (13)	**0.9567(8)**	1.2834 (5)	1.0478 (4)
	CS	1.1479 (13)	0.9317(6)	**0.7820 (2)**	0.9314 (3)
Iris	DBI	0.1099 (4)	0.1687 (3)	0.1011 (4)	**0.0998(4)**
	CS	2.7777 (4)	2.8412 (3)	2.5013(4)	**1.3783(4)**

The proposed method shows the best lowest clustering results and outperforms in a few prior researches [20], [21] and [22]. As indicated in Table 7, all existing methods have their complexity using evolutionary strategy automatic clustering methods for determining the number of clusters automatically in k-means, while the proposed method uses technical-statistical control for problem-solving of ACDE as an automatic clustering strategy for finding the optimal number of clusters in k-means method.

5 Conclusions

In this study, the proposed method is employed to find out the number of clusters in the k- means method. The tests were carried out using six datasets of UCI repository. Then, there is a comparison of the proposed method with previous related studies of k-Means, ACDE-k-means, GCUK, and DCPSO. The five methods are measured by DBI and CS evaluations. The method with the smallest evaluation value of the two evaluations that approaches zero is the best method. The experimental result reveals that the proposed method has the lowest value both in DBI and CS measure for all datasets compared with prior researches. These results showed that the proposed method achieves an excellent and promising performance. The use of U-Control Chart (UCC) method with automatic clustering differential evolution (ACDE) to determine the number of clusters in k-means has been proven to increase the performance of ACDE. Hence, it can be concluded that the U-Control Chart method can enhance ACDE-k-means performance in determining the number of clusters in k-means.

References

1. Ben Salem, S., Naouali, S., Chtourou, Z., 2018. A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach. Comput. Electr. Eng. 68, 463–483. https://doi.org/10.1016/j.compeleceng.2018.04.023.
2. Chakraborty, S., Das, S., 2018. Simultaneous variable weighting and determining the number of clusters—A weighted Gaussian means algorithm. Stat. Probab. Lett. 137, 148– 156. https://doi.org/10.1016/j.spl.2018.01.015.
3. Abdul Masud, M., Zhexue Huang, J., Wei, C., Wang, J., Khan, I., Zhong, M., 2018. Inice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres. Inf. Sci. (Ny). https://doi.org/10.1016/j.ins.2018.07.034
4. Rahman, M.A., Islam, M.Z., Bossomaier, T., 2015. ModEx and Seed-Detective: Two novel techniques for high quality clustering by using good initial seeds in K-Means. J. King Saud Univ. - Comput. Inf. Sci. 27, 113–128. https://doi.org/10.1016/j.jksuci.2014.04.002
5. Rahman, M.A., Islam, M.Z., 2014. A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowledge-Based Syst. 71, 345–365. https://doi.org/10.1016/j.knosys.2014.08.011
6. Ramadas, M., Abraham, A., Kumar, S., 2016. FSDE-Forced Strategy Differential Evolution used for data clustering. J. King Saud Univ. - Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2016.12.005
7. Yaqian, Z., Chai, Q.H., Boon, G.W., 2017. Curvature-based method for determining the number of clusters. Inf. Sci. (Ny). https://doi.org/10.1016/j.ins.2017.05.024

8. Tîrnăucă, C., Gómez-Pérez, D., Balcázar, J.L., Montaña, J.L., 2018. Global optimality in k-means clustering. Inf. Sci. (Ny). 439–440, 79–94. https://doi.org/10.1016/j.ins.2018.02.001
9. Xiang, W., Zhu, N., Ma, S., Meng, X., An, M., 2015. A dynamic shuffled differential evolution algorithm for data clustering. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.01.058
10. Garcia, A.J., Flores, W.G., 2016. Automatic Clustering Using Nature-Inspired Metaheuristics: A Survey. Appl. Soft Comput. https://doi.org/10.1016/j.asoc.2015.12.001
11. Das, S., Abraham, A., Konar, A., 2008. Automatic Clustering Using an Improved Differential Evolution Algorithm. IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans 38, 218–237. https://doi.org/10.1109/TSMCA.2007.909595
12. Omran, M.G.H., Engelbrecht, A.P., Salman, A., 2006. Dynamic clustering using particle swarm optimization with application in image segmentation. Pattern Anal. Appl. 332–344. https://doi.org/10.1007/s10044-005-0015-5.
13. Bandyopadhyay, S., Maulik, U., 2002. Genetic clustering for automatic evolution of clusters and application to image classification. Pattern Recognit. 35, 1197–1208.
14. Tam, H., Ng, S., Lui, A.K., Leung, M., 2017. Improved Activation Schema on Automatic Clustering Using Differential Evolution Algorithm. IEEE Congr. Evol. Comput. 1749–1756. https://doi.org/10.1109/CEC.2017.7969513
15. Kuo, R.., Suryani Erma, Yasid, A., 2013. Automatic Clustering Combining Differential Evolution Algorithm and k-Means Algorithm. Proc. Inst. Ind. Eng. Asian Conf. 2013 1207–1215. https://doi.org/10.1007/978-981-4451-98-7
16. Piotrowski, A.P., 2017. Review of Differential Evolution population size. Swarm Evol. Comput. 32, 1–24. https://doi.org/10.1016/j.swevo.2016.05.003
17. Kaya, I., 2009. A genetic algorithm approach to determine the sample size for attribute control charts. Inf. Sci. (Ny). 179, 1552–1566. https://doi.org/10.1016/j.ins.2008.09.024
18. Dobbie, G., Sing, Y., Riddle, P., Ur, S., 2014. Research on particle swarm optimization based clustering : A systematic review of literature and techniques. Swarm Evol. Comput. 17, 1–13. https://doi.org/10.1016/j.swevo.2014.02.001
19. Departamento Administrativo Nacional de Estadística. (2018). Página principal. Recuperado de:DANE http://www.dane.gov.co/
20. Torres-Samuel M., Vásquez C.L., Viloria A., Varela N., Hernández-Fernandez L., Portillo-Medina R. (2018) Analysis of Patterns in the University World Rankings Webometrics, Shanghai, QS and SIR-SCimago: Case Latin America. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.
21. Vásquez, C., Torres, M., Viloria, A.: Public policies in science and technology in Latin American countries with universities in the top 100 of web ranking. J. Eng. Appl. Sci. 12(11), 2963–2965 (2017)
22. Torres-Samuel M, Carmen Vásquez, Amelec Viloria, Tito Crissien Borrero, Noel Varela, Danelys Cabrera, Mercedes Gaitán-Angulo, Jenny-Paola Lis-Gutiérrez. (2018). Efficiency Analysis of the Visibility of Latin American Universities and Their Impact on the Ranking Web. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham