DOI: 10.24996/ijs.2022.63.12.41

ISSN: 0067-2904

Intelligent Bat Algorithm for Finding Eps Parameter of DbScan Clustering Algorithm

Khalil Ibrahim Ghathwan*, Athraa Jasim Mohammed

Computer Science Department, University of Technology , Iraq, Baghdad, Iraq

Received: 17/11/2021 Accepted: 22/2/2022 Published: 30/12/2022

Abstract

Clustering is an unsupervised learning method that classified data according to similarity probabilities. DBScan as a high-quality algorithm has been introduced for clustering spatial data due to its ability to remove noise (outlier) and constructing arbitrarily shapes. However, it has a problem in determining a suitable value of Eps parameter. This paper proposes a new clustering method, termed as DBScanBAT, that it optimizes DBScan algorithm by BAT algorithm. The proposed method automatically sets the DBScan parameters (Eps) and finds the optimal value for it. The results of the proposed DBScanBAT automatically generates near original number of clusters better than DBScanPSO and original DBScan. Furthermore, the proposed method has the ability to generate high quality clusters with minimum entropy [0.2752, 0.4291] in TR11 and TR12 datasets.

Keywords: BAT algorithm, DBScan, Eps parameter.

خوارزمية الخفاش الذكية لايجاد معامل خوارزمية كثافة البيانات الموجودة في المجموعة

خليل ابراهيم غثوان* , عذراء جاسم محمد

قسم علوم الحاسوب، الجامعة التكنولوجية، بغداد، العراق

الخلاصة

التجميع هو طريقة تعلم غير خاضعة للإشراف تصنف البيانات وفقًا لاحتمالات التشابه. DBScanهي خوارزمية عالية الجودة تم تقديمها لتجميع البيانات المكانية نظرًا لقدرتها على إزالة الضوضاء (الخارجة) وإنشاء أشكال عشوائية. ومع ذلك، فإنه يواجه مشكلة في تحديد القيمة المناسبة لمعاملEsp . يقترح في هذا البحث طريقة تجميع جديدة، مصطلحات مثل DBScanBAT ، والتي تعمل على تحسين خوارزمية DBScan بواسطة خوارزميةBAT . الطريقة المقترحة تقوم تلقائيًا بتعيين معلمات (Eps) DBScan والعثور على القيمة المثلى لها. تولد نتائج DBScanBAT المقترحة تلقائيًا عددًا قريبًا من المجموعة الأصلية أفضل من DBScanPSOو DBScan الأصلي. علاوة على ذلك، فإن الطريقة المقترحة لديها القدرة على إنشاء مجموعات عالية الجودة مع الحد الأدنى من الانتروبيا [0.2752 ، 0.4291] في مجموعات بيانات TR11 وTR12 .

*Email: 110039@uotechnology.edu.iq

1. Introduction

Data clustering is a familiar data mining method that provides affective result in many areas. It is the grouping of the objects in the datasets into clusters according to the similarity between these objects. Many clustering algorithms have been proposed in literature and these algorithms are classified into supervised and unsupervised methods. A supervised method needs to determine the number of the clusters as initial value. Example of this type is K-means [1]. While unsupervised methods are not required to specify the number of the clusters at the start, for example, Density Based Spatial Clustering of Application with noise (DBScan) [2].

DBScan was proposed by Ester et al. [2, 3] and is a well-known clustering algorithm in scientific literature. It requires two important parameters as initial values; the radius of a neighborhood (Eps) and the minimum number of points required for constructing a dense cluster (MinPts). DBScan operates by determining the starting point not visited previously. The neighborhood of this point is collected and if it includes a large number of points, a cluster is initiated; otherwise, this point is classified as noise point [4]. DBScan has significant advantages such as its ability to discover clusters; it does not need to determine the number of clusters and being able to process huge datasets. However, the original DBScan has difficulty in determining the optimal radius of a neighborhood (Eps) value. Hence, a modification on original DBScan is needed to automatically obtain the optimal Eps [5]. The pseudo code for DBScan is shown in Figure 1.

```
- Input dataset (DS), distance function (D),
  the radius of a neighborhood (eps),
  the minimum number of points (minPts).
- Initiate number of cluster C=0.
  Step1: For i= 1 to n (n size of dataset) {
  Step2:     Neighbors N = RangeQuery(DS, D, i, eps)
  Step3:     If N < minPts then
  Step4:         Label(i) = noise }
  Step5:         C=C+1
  Step6:         Label(i) = C
  Step7:         Seed set S = N \ {i}
  Step8:         for each point j in S {
  Step9:                 if label(j) = Noise then label(j) = C
  Step10:                        Neighbors N = RangeQuery(DS, D, j, eps)
  Step11:                        if |N| ≥ minPts then {
  Step12:                                S = S ∪ N     }   }   }
```

Figure 1: Pseudo Code of Density Based Spatial Clustering (DBScan) [2].

In this paper, a new method is proposed to optimize DBScan using the Bat algorithm to automatically set the optimal Eps value with different datasets. Bat Algorithm (BA) is a well-known meta-heuristic algorithm. It was first introduced by Xin-Shin Yang [6]. In previous studies [7, 8, 9, and 10], the Bat algorithm was proven successful in finding the global optimization value. Bat is using echolocation to sense distance. Every bat has its own position (Xj), a velocity (Vj), a loudness (A), a pulse rate (r), and a frequency (Fj) that exemplify the objective function. When the Bat searches for its prey, it moves randomly and changes its position, velocity, and frequency. The following equations (1, 2, and 3) show how to calculate the new position, new velocity, and new frequency [6].

$$Fj=Fmin+(Fmax-Fmin)\ \beta \tag{1}$$

$$Vjt=Vjt\text{-}1 + (Xjt - X^*)\ Fj \tag{2}$$

$$Xjt= Xjt\text{-}1 + Vjt \tag{3}$$

Where: β is a random number between 0 and 1. X* is best position (best solution) among all bats. The Pseudo-code of Bat algorithm is shown in Figure 2.

```
Input:
- Determine the number of Bats (n).
- Randomly, determine the initialized position ($X_n$).
- Randomly, determine the initialized velocity ($V_n$).
- Determine the frequencies Fn, pulse rate r, loudness A.
Process:
- Evaluate each Bat and find best.
  Step1: While not stopping criterion do
  Step2:      For j = 1 to n do
  Step3:              Generate new solution by adjusting frequency,
                      update velocity,
                      and update position using Equations 1, 2, and 3
  Step4:                      If (rand > r)
  Step5:                              Generate a solution around the best solution
  Step6:                       End if
  Step7:                      If (rand<A & F(Xj) < F(X*))
  Step8:                              Replace the current solution with new solution
  Step9:                              Replace the current fitness with new fitness
  Step10:             End if
  Step11: End while
Output:
- Return the best solution.
```

Figure 2: Pseudo Code of basic Bat algorithm [6].

The rest of the paper is organized as follows: Section 2 includes a brief explanation of related work. Section 3 introduces the proposed method. The results are presented in Section 4, while the conclusion is given in Section 5.

2. Related Work

The density based clustering algorithm is depended on “the idea that objects which form a dense region should be grouped into one cluster” [2]. It utilizes a fixed threshold value to distinguish high density regions from low density regions.

DBScan has considerable benefit in the clustering such as not requiring a pre-determination of the number of clusters, and dealings with large datasets. Despite of this advantage, DBScan still needs parameters setting the same as other clustering algorithm [11]. Various values of DBScan parameters (Eps and MinPts) will produce different clustering results in the same dataset. Further, in literature there does not exist any theoretical guidance for setting the parameters (Eps and MinPts), where the selection of DBScan parameters is usually based on many experimental attempts (i.e., trials). Since the performance of DBScan clustering is sensitive to the parameters (Eps and MinPts), it is required to set them previously.

To address this problem, different studies were conducted previously. Lai et al. [11] proposed improved MVO algorithm to optimize DBScan for searching the range of Eps. Rahmah, and Sitanggang [5] determination of Eps value in DBSCAN algorithm was also a solution suggested, where the researchers modified the DBscan algorithm using Euclidean distance for each pair of data. Then, the k-nearest neighbor search was used to determine the optimal Eps value. The researchers used a search algorithm that finds the minimum values of distance to nearest three and sort distance ascending and plot each value to find Eps that corresponds to critical changes in a carve.

Ozkok and Celik [12] and, Wang et al. [13] proposed to use k-list to automatically determine Eps' value. Birant, and Kut [2] present improved DBSCAN that can detect noise in various density data, can cluster spatial data, and can cluster the border objects. Shen et al. [14] introduced a time super pixel algorithm using DBScan, where the DBSCAN was employed to cluster pixels with color similarity. Due to its outperformance, it is considered the state-of the art method in accuracy and efficiency.

Tran, Drab, and Daszykowski [15] proposed a modification of the DBSCAN algorithm to detect the border object of adjustment clusters. The revised DBSCAN outperformed the original algorithm. Chehreghani, Abolhassani, and Chehreghani [4] suggested an improvement to DBSCAN for hierarchal clustering of web pages. It was based on three phases; the distance calculation, the extraction stage (where each distance is examined to find the effectiveness on the clustering), and the combination stage, where the researchers used improved single and average linkage method. The results show good generated clusters with high quality.

In [16], Karami and Johansson proposed to utilize a binary differential evolution algorithm to optimize the DBScan parameters (Eps) and find the interval of Eps. Guan et al. [17] introduced a new method that's based on using Particle Swarm Optimization (PSO) with DBScan to solve the problem of a difficult set of two critical parameters of DBScan. Latifi-Pakdehi and Daneshpour [18] proposed a hierarchical clustering to generate a number of values to improve the Eps parameter. Valarmathy, and Krishnaveni [19] used K-distance graph method to develop a new model to select Eps and MinPts automatically as well as to increase the speed of the process.

To our knowledge, there is no method in previous studies that utilize Bat algorithm to optimize the parameters of DBScan. The improved Harmony search algorithm (HS) was also used to optimize DBSCAN, when KDBSCAN combined with HS to optimize the clustering parameters Eps and MinPts [20]. Chin et al. [21], proposed the use of DBSCAN as a Self-Adaptive DBSCAN-based method for solving a problem with wafer bin map such as the noise point detection of wafer maps (DBSCANWBM).

3. Proposed Method (DBScanBAT)

This paper introduces a new clustering method based on DBScan and Bat algorithm, called DBScanBAT, to automatically discover the optimal value of Eps. Figure 3 shows the process of the proposed DBScanBAT

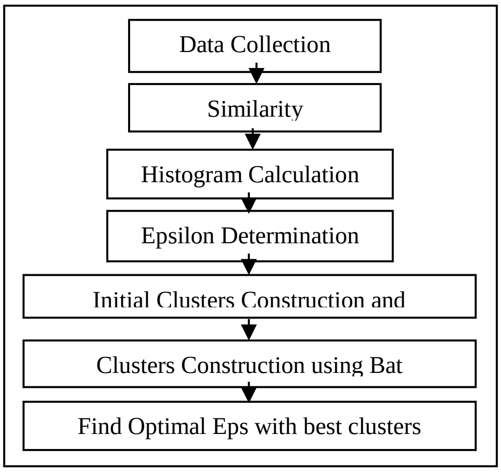

Data Collection
Similarity
Histogram Calculation
Epsilon Determination
Initial Clusters Construction and
Clusters Construction using Bat
Find Optimal Eps with best clusters

Figure 3: The process of the proposed DBScanBAT.

3.1 Data Collection

In this paper, benchmark datasets from TREC collection from Cluto toolkit [22], named TR11 and TR12 were used (shown in table 1). In these two datasets, the TR11 dataset contained 414 web pages and nine clusters with 6429 words. The TR12 dataset contained 313 web pages, eight classes and the number of words was 5804. The maximum number of webs in a class was 132 in TR11 dataset and 93 in TR12 dataset, While the Minimum number of webs in class 6 in TR11 and 9 in TR12.

Table 1: Description of TR11and TR12 datasets

Dataset	Number of web pages	Number of clusters (classes)	Minimum number of webs in class	Maximum number of webs in class	Number of words
TR11	414	9	6	132	6429
TR12	313	8	9	93	5804

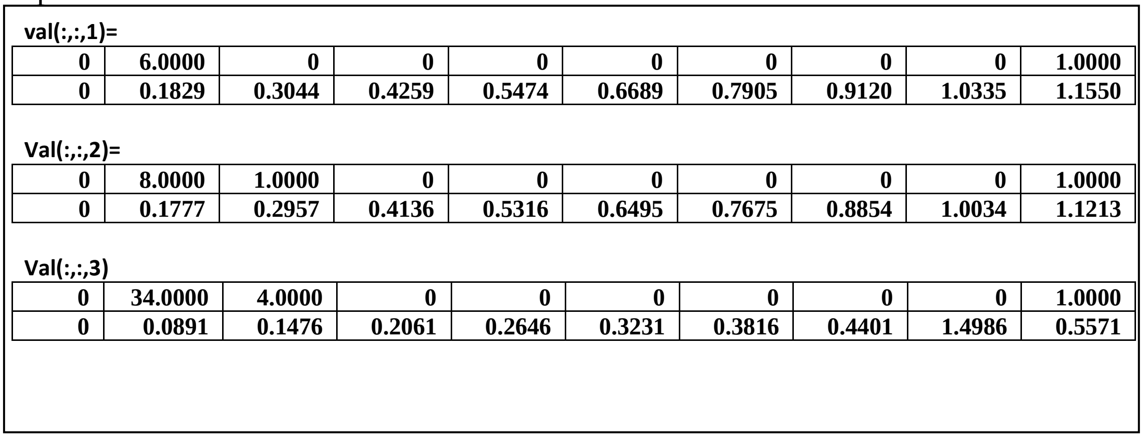
3.2 Similarity Calculation

The next step in the process of DBScanBAT is a similarity calculation between the web pages. The cosine similarity is applied to measure the similarity [23]. The cosine similarity is shown in equation 4.

$$\text{Cosine} - \text{Similarity}\ (\text{Di}, \text{Dj}) = \sum_{i=1,j=1}^{m} (\text{Di} * \text{Dj}) \qquad (4)$$

3.3 Histogram Calculation

The third step in the proposed DBScanBAT algorithm is the histogram calculation, where, each row of dataset of web pages bins the element into 10 equally spaced containers and returns the number of elements in each container. Figure 4 shows the matrix created after this step.



val(:,:,1)=

0	6.0000	0	0	0	0	0	0	0	1.0000
0	0.1829	0.3044	0.4259	0.5474	0.6689	0.7905	0.9120	1.0335	1.1550

Val(:,:,2)=

0	8.0000	1.0000	0	0	0	0	0	0	1.0000
0	0.1777	0.2957	0.4136	0.5316	0.6495	0.7675	0.8854	1.0034	1.1213

Val(:,:,3)

0	34.0000	4.0000	0	0	0	0	0	0	1.0000
0	0.0891	0.1476	0.2061	0.2646	0.3231	0.3816	0.4401	1.4986	0.5571

Figure 4: The matrix of histogram

3.4 Epsilon Determination

The initial parameter values (Eps) of DBScan be determined based on the matrix of histogram, where, the value that has the maximum number of elements in each container is assigned to Eps. Therefore, a matrix of Eps is constructed.

3.5 Initial Clusters Construction and Evaluation

In this step, from each value in the matrix of Eps, the clusters are constructed using DBScan algorithm. The minimum number of points required for constructing a dense cluster (MinPts) is set to (10). After that, the created clusters are evaluated using the performance metric entropy [24], shown in equation 5.

$$\text{Entropy} = \sum_{i=1}^{k} \frac{\text{Hj} * |\text{Cluster}|}{\text{M}} \qquad (5)$$

$$\text{Hj} = -\sum_{k=1}^{CN} \frac{|\varphi_k \cap C_j|}{|C_j|} \log \frac{|\varphi_k \cap C_j|}{|C_j|} \qquad (6)$$

Where, K means number of clusters, CN output cluster, $\varphi_k$is known class.

3.6 Clusters Construction using Bat Algorithm

In this step, the Bat algorithm is used. The entropy values that were collected from the previous step are assigned as frequencies of Bat algorithm and the best Eps that generated clusters of lower entropy is assigned as best position of BAT algorithm. The Pseudo code of the proposed DBScan with BAT algorithm is shown in Figure 5.

```
Input:
- Determine the number of Bats (n).
- Determine the initialized position (X_n) which equal to matrix of Eps.
- Randomly, determine the initialized velocity (V_n).
- Determine the frequencies Fn which equal to matrix of Entropy.
- Determine the pulse rate r, loudness A.
Process:
  Evaluate each Bat and find best.
  Step1:  While not stopping criterion do
  Step2:     For j = 1 to n do
  Step3:       Generate new solution by adjusting frequency,
               update velocity,
                 update position using Equations 1, 2, and 3
  Step4:             If (rand > r)
  Step5:                        Generate a solution around the best solution
  Step6:             End if
  Step7:             If (rand<A & F(Xj) < F(X*))
  Step8:                        Replace the current solution with new solution
                                Replace the current fitness with new fitness
  Step9:             End if
  Step10:End while
Output:
- Return the best solution (best Eps with lower Entropy).
```

Figure 5: The Pseudo code of proposed DBScan with BAT algorithm

4. The Results

In this paper, the proposed DBScanBAT method is evaluated by comparing it to the DBScanPSO and the original DBScan algorithm using entropy metric. The parameters setting of the Bat and PSO algorithm are presented in Table 2.

Table 2: Parameters setting of Bat and PSO algorithms

Bat algorithm Parameters Names	Setting Values	PSO algorithm Parameters Names	Setting Values
Number of iterations	20	Number of iterations	20
A	0.25	W	0.72
r	0.5	C1, C2	1.49

Experiments were conducted in MATLAB R2018b. Table 3 demonstrates the Eps value and entropy results of all methods; DBScanBAT, DBScanPSO and DBScan.

As shown in Table 3 and Figure 6, in TR11 dataset, DBScanBAT algorithm automatically generates five clusters, Eps value is (0.1793) with lower entropy value (0.2752). While, DBScanPSO algorithm generates a number of clusters, i.e., 5, Eps value is (0.1868) with lower entropy (0.2752). DBScan algorithm generates several clusters (3) far from the original dataset, Eps value is (0.1744) with an entropy value (0.8920). Concluding from this, DBScanBAT algorithm has better entropy compared with the original DBScan algorithm.

Table 3: The Results of DbScanBat, DbScanPSO and DbScan Algorithms

Dataset	Methods	No. of cluster	Best Eps	Entropy
TR11	DBScanBAT	5	0.1793	0.2752
	DBScanPSO	5	0.1868	0.2752
	DBScan	3	0.1744	0.8920
TR12	DBScanBAT	4	0.1803	0.4291
	DBScanPSO	3	0.1661	0.4140
	DBScan	4	0.1700	0.7936

In addition, in TR12 dataset and Figure 6, DBScanBAT algorithm automatically produces a better number of clusters (4) compared to DBScanPSO algorithm that generates (3). Despite of the entropy of DBScanPSO (0.4140) being lower than proposed DBScanBAT. Also, the Eps value is (0.1803) generated by proposing DBScanBAT algorithm with smaller entropy value (0.4291) compared with the entropy value of DBScan algorithm which is (0.7936).

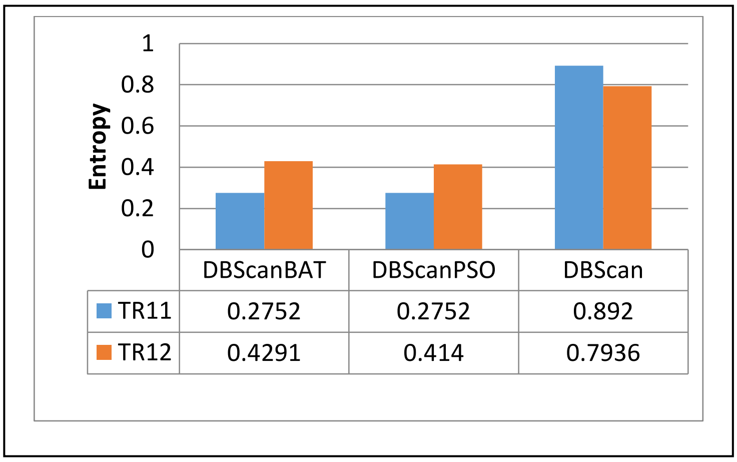


Figure 6: Entropy values of DBScanBAT, DBScanPSO and DBScan algorithms

It is observed that the proposed DBScanBAT algorithm generates best Entropy value. Lower entropy means best clustering result. Therefore, the best Eps value is generated by DBScanBAT algorithm.

5. Conclusion

In this study, a new method is proposed by combining Bat algorithm with DBScan clustering method. The aim of this paper was to determine the best parameter of DBScan (Eps) that can lead to generating good clusters with high quality. Previous studies demonstrated that Bat algorithm has the ability to find the optimal solution. The proposed method firstly fond the Eps value using the histogram, then created the clusters and evaluated these clusters using entropy.

The evaluation values were assigned to Bat algorithm as the initial value of frequencies. DBScanBAT algorithm has better entropy compared with the original DBScan algorithm. Experimental results demonstrate that the proposed method is given best result compareed with the original DBScan algorithm. Moreover, the results show the proposed method has the ability to create high quality clusters with minimum entropy [0.2752, 0.4291] in TR11 and TR12 datasets.

6. References

[1] A. K. Jain,“Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, Issue 8, pp. 651-666, 2010.

[2] D. Birant, A. Kut, “ST-DBSCAN: An algorithm for clustering spatial–temporal data,” *Data & Knowledge Engineering*, vol. 60, Issue 1, pp. 208-221, 2007.

[3] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise”, in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, pp. 226–231, 1996.

[4] M. H. Chehreghani, H. Abolhassani, M. H. Chehreghani, “Improving density-based methods for hierarchical clustering of web pages,” *Data & Knowledge Engineering*, vol. 67, Issue 1, pp. 30-50, 2008.

[5] N. Rahmah, I. S. Sitanggang, “Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra,” *IOP Conf. Series: Earth and Environmental Science*, vol. 31, 2016.

[6] X. S. Yang, X. He, “Bat algorithm: literature review and applications,” *International Journal of Bio-Inspired Computation*. vol. 5, Issue 3, 2013.

[7] S. Srivastava, S. K. Sahana, “Application of Bat Algorithm for Transport Network Design Problem,”*Applied Computational Intelligence and Soft Computing*, vol. 2019, Article ID 9864090, 12 pages, 2019.

[8] M. A. Al-Betar, M. A. Awadallah, “Island bat algorithm for optimization,”Expert Systems with Applications, vol. 107, pp. 126-145, ISSN 0957-4174, 2018.

[9] S. Nandy, P.P. Sarkar, “Chapter 8 - Bat algorithm–based automatic clustering method and its application in image processing,” Editor(s): Xin-She Yang, João Paulo Papa, Bio-Inspired Computation and Applications in Image Processing, Academic Press, pp. 157-185, ISBN 9780128045367, 2016.

[10] G. Wang, B. Chang, Z. Zhang, “A multi-swarm bat algorithm for global optimization,” *IEEE Congress on Evolutionary Computation (CEC)*, Sendai, pp. 480-485, 2015.

[11] W Lai, M. Zhou, F. Hu, K. Bian, Q. Song, “A new DBSCAN parameters determination method based on improved MVO,” *IEEE Access*, vol.7, 104085 – 104095, 2019.

[12] F. O. Ozkok and M. Celik, “A new approach to determine Eps parameter of DBSCAN algorithm”, Int. J. Intell. Syst. Appl. Eng., vol. 5, no. 4, pp. 247-251, 2017.

[13] W.-T. Wang, Y.-L. Wu, C.-Y. Tang and M.-K. Hor, “Adaptive density-based spatial clustering of applications with noise (DBSCAN) according to data”, Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC), pp. 445-451, Jul. 2015.

[14] J. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, and L. Shao, “Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm,” *IEEE Transactions on Image Processing, vol.* 25, Issue 12, pp. 5933 – 5942, 2016.

[15] T. N. Tran, K. Drab, M. Daszykowski, “Revised DBSCAN algorithm to cluster data with dense adjacent clusters,” *Chemometrics and Intelligent Laboratory Systems*, vol. 120, pp. 92-96, 2013.

[16] A. Karami and R. Johansson, "Choosing DBSCAN parameters automatically using differential evolution", *Int. J. Comput. Appl.*, vol. 91, no. 7, pp. 1-11, Apr. 2014.

[17] C. Guan, K. Kam, F. Yuen, F. Coenen,” Particle swarm Optimized Density-based Clustering and Classification: Supervised and unsupervised learning approaches,” *Swarm and Evolutionary Computation*, Vol. 44, pp.876-896, 2019.

[18] A. Latifi-Pakdehi, N. Daneshpour, "DBHC: A DBSCAN-based hierarchical clustering algorithm", *Data & Knowledge Engineering*, Vol. 135, pp.101922, 2021.

[19] N. Valarmathy, S. Krishnaveni, "A novel method to enhance the performance evaluation of DBSCAN clustering algorithm using different distinguished metrics", in: Proceedings of Materials Today, 2020.

[20] Q. Zhu, X. Tang, A. Elahi, "Application of the novel harmony search optimization algorithm for DBSCAN clustering", Expert Systems with Applications, Vol. 178, pp.115054, 2021.

[21] S. Chen, M. Yi, Y. Zhang, X. Hou, Y. Shang, P. Yang, "A self-adaptive DBSCAN-based method for wafer bin map defect pattern classification", Microelectronics Reliability, Vol. 123, pp. 114183, 2021.

[22] CLUTO - Software for Clustering High-Dimensional Datasets, http://glaros. dtc.umn.edu/gkhome /cluto/cluto/download

[23] C. Luo, Y. Li, S. M. Chung, "Text document clustering based on neighbors," *Data & Knowledge Engineering*, Vol. 68, Issue 11, pp. 1271-1288, 2009.

[24] K. Murugesan, J. Zhang, "Hybrid hierarchical clustering: An experimental analysis," university of Kentucky, 2011.