University of Abdelhamid Ibn Badis
Mostaganem, Algeria

Faculty of Sciences & Technology

8th – 9th
May 2022

Mostaganem
Algeria

The IEEE Seventh International Conference on Image and Signal Processing and their Applications

IEEE
Advancing Technology for Humanity

IEEE Algeria Section

CERTIFICATE OF PARTICIPATION

This is to certify that

Aida CHEFROUR

has presented a communication at the ISPA 2022 entitled:

'A Modified Incremental Density Based Clustering Algorithm'

IEEE
IEEE Xplore

Prof. M... Kamel
Chair ISPA 2022

IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

A modified Incremental Density Based Clustering Algorithm

Aida Chefrour

2022 7th International Conference on Image and Signal Processing and their Applications (ISPA)

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

You have indicated that you DO NOT wish to have video/audio recordings made of your conference presentation under terms and conditions set forth in "Consent and Release."

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Aida CHEFROUR	22-03-2022
Signature	**Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEEs publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.
Please direct all questions about IEEE copyright policy to:
IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966

2022 7th International Conference on Image and Signal Processing and their Applications (ISPA 2022)

Mostaganem, Algeria
May 8 – 9, 2022

Proceedings of ISPA 2022

Organized by:

Laboratory Signals & Systems

Sponsored by:

University of Abdelhamid Ibn Badis of Mostaganem, Algeria
Faculty of Sciences and Technology

Technical co-sponsors

2022 7th International Conference on Image and Signal Processing and their Applications (ISPA)

IEEE Catalog Number CFP22ONC-ART

ISBN 978-1-6654-8042-0

ISSN

Additional copies of this publication are available from

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA

+1 845 758 0400
+1 845 758 2633 (FAX)
email: curran@proceedings.com

Table of Contents

A modified Incremental Density Based Clustering Algorithm

Aida CHEFROUR
Computer Science Department
Mohamed Cherif Messaadia University
Souk Ahras, Algeria
LISCO Laboratory
Badji Mokhtar University
Annaba, Algeria
aida.chefrour@univ-soukahras.dz

***Abstract*—Cluster analysis, generally known as clustering, is a technique for separating data into groups (clusters) of similar objects. Except if the system is completely retrained, traditional clustering classifiers will be unable to learn new information and knowledge (attributes, examples, or classes). Only incremental learning, which outperforms when new data objects are introduced into an existing database, can solve this problem. These evolutionary strategies are applied to dynamic databases by updating the data. We'll choose to study the Incremental Density-Based Spatial Clustering of Applications with Noise algorithm because of its capacity to discover arbitrary clusters and identify noise. In this study, a modified version of the Incremental Density Based Clustering Algorithm using an Adaptive Median Filtering Technique was used. The difference between our previous proposed AMF-IDBSCAN and the proposed algorithm developed in this work is in the evaluation performance stage. The key idea consists of a database change in the case of introducing new data items to an existing database in order to improve performance. We conducted several experiments on benchmark and synthetic data collected from the University of California Irvine repository in terms of the Generalized Dunn Index, Davies Bouldin Index, and change of time (milliseconds) with the increment of data in the original database. Experiments with datasets of various sizes and dimensions show that the proposed algorithm enhances clustering when compared to several current incremental well-known techniques.**

***Index Terms*—Incremental clustering, DBSCAN, Adaptive median filtering, AMF-IDBSCAN**

I. Introduction

Data mining is a multidisciplinary field that has a variety of definitions [1]. Data analysis is primarily concerned with several large data repositories in the database management business, andt aims to find real, valuable, novel, and understandable patterns in existing data.

In data mining, clustering is a common data discovery approach. It divides a data set into subsets or clusters with similar traits or properties [2]. Its goal is to partition the data into groups (clusters) of objects that are comparable [3]. Items in the same cluster are more similar to one another than objects in other clusters. Clustering is frequently utilized in domains such as artificial intelligence, pattern recognition, statistics, and other data processing.

978-1-6654-8042-0/22/$31.00 ©2022 IEEE

There have been numerous clustering algorithms developed, which can be classified into the following basic types [4] hierarchical clustering methods (BIRCH, CHAMELEON, etc.), partitioning algorithms (K-means, K-medoids), density-based algorithms (DBSCAN, OPTICS), and grid-based algorithms (STING, CLIQUE). The distinction between traditional clustering methods (batch mode) and incremental clustering methods is that incremental clustering's capacity to process new data added to the data collection without requiring a full re-clustering. During clustering, this enables dynamic tracking of database updates. Incremental learning has received a great deal of press in recent years because it provides for effective data reuse, quick and pragmatic learning based on context, knowledge augmentation, learning in dynamic and big databases, exploration, and smart decision-making [5]. We are interested in evolving incremental clustering to cluster data items, which is the procedure of incrementally upgrading an existing collection of clusters rather than mining them from scratch with each database update [6]. Evolving clustering algorithms enable gradual structural and parametric modifications to be made using various data-driven processes [7]. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm is the subject of our research. DBSCAN's core concept is that each object in a cluster must be surrounded by a specific number of other objects. When compared to other clustering techniques, this one is superior. DBSCAN offers a number of best features, and it's been utilized by a great number of researchers in recent years due to its various extensions and applications.

We were particularly interested in the incremental version of DBSCAN because (1) it can find clusters of any shape; (2) it only requires two parameters and is unconcerned about the ordering of the points in the database; (3) it reduces the search space and allows for incremental updates in the clusters [8]; and (4) it is more adaptable to various datasets and data space without requiring any special software [9].

In this paper, we propose a modified version of the AMF-IDBSCAN algorithm developed by [10]. We remember the principle of the AMF-IDBSCAN: our algorithm is divided into three parts. It preprocesses the original database after importing it to prepare for the clustering phase and minimize the dataset's volume using Canopy clustering. The findings of the first stage are then applied to the results of the DBSCAN algorithm to create a new database. The incremental dataset is then subjected to the incremental DBSCAN algorithm. To reduce noise and outliers, the adaptive median filtering approach is used to incorporate the findings of the preceding phase. The results are then compared, and the overall performance is evaluated.

The evaluation performance stage compares the efficiency of the modified evolving AMF-IDBSCAN algorithm to static DBSCAN, and incremental DBSCAN of [11],in terms of the Dunn Index and Davies-Bouldin criteria, as well as the change in time (milliseconds) with the increment of data in the original database. In this last one, the data increment is known as the percent delta change. We chose to compare these algorithms because they give reasonably different results, given that if we compare it with other.

The rest of the paper is organized as follows: In the next section, we survey in brief the literature of enhanced incremental DBSCAN algorithms. In section 3, we describe our proposed modified AMF-IDBSCAN algorithm. Section 4 describes the experiment we conducted and the results obtained by our algorithm. Finally, we draw some conclusions and show ongoing research aspects in Section 5.

II. Related Work

In the literature, there are several algorithms for improving DBSCAN. We'll go over the most well-known and recent ones in this part:

IDBSCAN [12] proposes an improved version of the incremental DBSCAN technique for constructing and updating arbitrary-formed clusters in large datasets incrementally. The proposed approach enhances incremental clustering by restricting the search space to partitions rather than the original dataset, resulting in significant performance improvements when compared to other incremental clustering algorithms.

[13] offers an incremental DBSCAN that is combined with a suitable noise removal and outlier detection strategy inspired by the box plot method to improve this algorithm. To frame the remaining number of clusters, it employed a network measure for dense regions.

An enhancement of the DBSCAN algorithm is proposed by [10] based on incremental clustering called AMF-IDBSCAN which builds incrementally the clusters of different shapes and sizes in large datasets and eliminates the presence of noise and outliers. The proposed AMF-IDBSCAN algorithm uses a canopy clustering algorithm to pre-clustering datasets to decrease the volume of data, applies an incremental DBSCAN for clustering the data points, and Adaptive Median Filtering (AMF) technique for post-clustering to reduce the number of outliers by replacing noise with chosen medians.

III. Proposed Modified AMF-IDBSCAN Clustering Algorithm

We remember the principle of our algorithm AMF-IDBSCAN, developed by Chefrour and Souici-Meslati [10]:

It is an upgraded incremental DBSCAN that uses a canopy clustering method and an adaptive median filtering technique to address the limitations of classic clustering algorithms, excessive complexity and non-scalability.

The proposed AMF-IDBSCAN is divided into four stages: Preclustering using Canopy clustering is the first step. The clustering of data items is the second phase, which employs incremental DBSCAN.

Post-clustering is the third phase, which uses the Adaptive Median Filtering approach to lower the number of outliers by replacing them with specified medians. In the last phase, multiple evaluation metrics are employed to assess the performance of clustering algorithms.

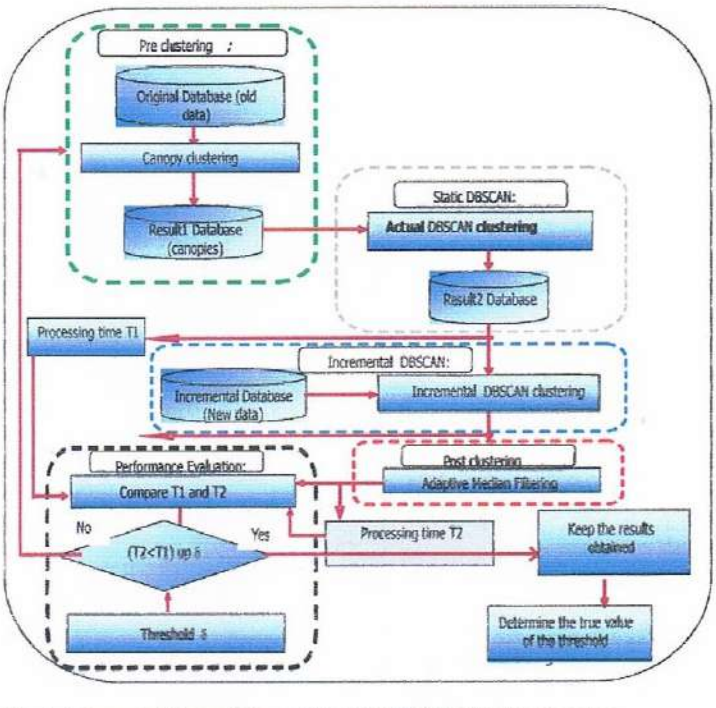

Fig. 1. The methodology of the modified AMF-IDBSCAN clustering algorithm.

The difference between the modified version proposed in this work and the AMF-IDBSCAN algorithm is:

- Determine the static DBSCAN execution time (T1) on the original database.
- Determine the time required to execute incremental DBSCAN and AMF techniques on the incremental database (T2).
- Compare (T1, T2):
  1) If $(T2 < T1)$ up to a certain threshold value in the database, then keep the results and determine the real threshold value;
  2) Otherwise, go to step 1 of the AMF-IDBSCAN algorithm (pre-clustering step).

Pseucode Algorithm

Input

D: a dataset containing n objects $X1, X2, X3\ldots, Xn$;
n: number of data items;
$Minpts$: Minimum number of data objects ;
eps: radius of the cluster;
CN: canopies centers.
Output
K: a Set of clusters;
A single value: $I_x, y, ...$ or $Imed$
Procedure
BEGIN

1) **Run Canopy clustering:**
   - Put all data into a List, and initialize two distance radius about the loose threshold $T1$ and the tight threshold $T2(T1 > T2)$.
   - Randomly select a point as the first initial center of the Canopy cluster, and delete this object from the List.
   - Get a point from the List, and calculate the distance d to each Canopy clusters.
     If $d < T2$, the point belongs to this cluster;
     if $T2 <= d <= T1$, this point will be marked with a weak label;
     If the distance d to all Canopy center is greater than $T1$, then the point will be classed as a new Canopy cluster center.
     Finally, this point should be eliminated from the List;
   - Run the step (Get a point from the List) repeatedly until the list is empty, and recalculate the canopy centers CN.
2) **Run the actual DBSCAN algorithm** and clustered the new data item C_i properly based on the radius(eps) and the Minpts criteria. Repeat till all data items are clustered:
   - Choosing Canopy Centers CN;
   - Attribute data points D to canopy centers CN;
   - Apply the DBSCAN algorithm with radius$\epsilon <=$ canopy radius with $dist(CN, C_i) < Minpts$;
   - Repeat the iteration until all data are clustered. **DBSCAN T1 (Processing Time).**
3) **Run the incremental DBSCAN:**
   - Let, K represents the already existing clusters.
   - When new data is coming into the database, the new data will be directly clustered by calculating the minimum $mean(M)$ between that data and the core objects of existing clusters.
     For $i = 1$ to n do
     find some $meanM$ in some cluster K_p in K such that $dis(C_i, M)$ is the smallest;
     If $(dis(C_i, M)$ is minimum) and$(C_i <= eps)$ and $(size(K_p) >= Minpts)$ then
     $K_p = K_pUC_i$;
     Else
     If $dis(C_i! = min)||(C_i > eps)||(size(K_p) < Minpts)$ then
     C_i take Outlier(O_i).
   - Elimination of noise objects O_i:
     The new dataset contains O_i and the clusters closest to it.
4) **Adaptive Median Filtering Technique:**
   - For $i = 1$ to m do where m is the number of outliers Illustrate a new rectangle on a hyperplane;
     Let:
     $I_x, y...$ be the selected noise data (O_i) at the coordinates ($(x, y, ...)$; corresponding the data dimensions)
     I_min be the minimum noise value;
     I_max be the maximum noise value in the window;
     W be the current window size applied; (It contains K clusters and O_i);
     W_max be the maximum window size that can be reached;
     I_med be the median of the window assigned
     Algorithm
     Level A: $A_1 = I_med - I_min$
     $A_2 = I_med - I_max$
     If $A_1 > 0$ AND $A_2 < 0$, Go to level B
     Else increase the window size
     If window size $W <= I_max$ repeat level A
     Else output $I_x, y...$
     Level B: $B_1 = I_x, y, .-I_min$
     $B_2 = I_x, y, ..-I_min$
     If $B_1 > 0$ And $B_2 < 0$ output $I_x, y, ..$
     Else output I_med.
   - Repeat step b till all the data samples are clustered. **Incremental DBSCAN T2 (Processing Time).**
5) **Evaluate performance:**
   - Compare $(T1, T2)$;
     If $(T2 < T1)$ up to a certain threshold value in the database then keep the same results else go to 1 (carry out the canopy clustering).

END.

IV. Experiments and Results

A. Performance metrics

Cluster validity refers to the process of evaluating the results of a clustering method. Cluster validity can be studied in three methods. The first is based on external criteria that assess the results of a clustering algorithm using a pre-specified structure that is placed on a dataset and reflects our knowledge about the data set's clustering structure. The second is based on internal criteria, in which the clustering results are evaluated using values that involve the data set's vectors. The third method of determining clustering validity is based on relative criteria, which involves comparing a clustering structure to alternative clustering schemes produced by the same algorithm but with different input parameter values [14].

We used relative criteria in our study since internal and external criteria have the disadvantage of requiring a great deal of computation [15].

Dunn's index vD: This index (see Eq. 1) is used to find compact and well-separated clusters:

$$vD = min_{i \in c} min_{j \in c, i \neq j} \gamma(c_i, c_j) / max_{k \in c} \delta(c \in k) \quad (1)$$

where

$\gamma(c_i, c_j) = min_{(d(x_i, x_j)) | x_i \in c_i, x_j \in c_j}$

$\delta(c_k) = max_{d(x_i, x_j)) | x_i, x_j \in c_i}$

d is a distance function and C_i, C_j, C_k are the sets whose elements are the data points assigned to the corresponding i_{th}, j_{th} and k_{th} clusters respectively. The main disadvantage of direct implementation of Dunn's index is that it is computationally expensive to calculate as the number of clusters and total points grows. The number of clusters that maximize vD is considered the optimal number of clusters [15]. Larger values of vD correlate to the right clusters, and the number of clusters that maximize vD is considered the best number of clusters.

Davies-Bouldin index: The ratio of within-cluster scatter to between-cluster separation (see Eq. 2) determines this index:

$$DBI = 1/n \sum_{i=1}^{n} max_{i \neq j}((S_n(Q_i) + S_n(Q_j))/(S(Q_i, Q_j)) \quad (2)$$

Where n is the number of clusters, S_n is the average distance between cluster centers, and $S(Q_i, Q_j)$ is the distance between cluster centers. As a result, if the clusters are compact and far apart, the ratio is low. For a decent cluster, the Davies-Bouldin index will have a tiny value [15].

B. Databases Description

We chose five data sets from the UCI machine learning repository to validate the accuracy and efficiency of our proposed method (3D Road Network, Wine, Glass identification, Letter recognition, and Fisher's Iris) (see Table. I). Three of them are small data sets, while the other two are considered large data. We compare the efficiency of the proposed modified

TABLE I
DESCRIPTION OF UCI DATABASES

Dataset	No. of instances	No. of attributes	Attribute type	Data types
3D Road Network	434874	4	Numerical	Real
Wine	178	13	Numerical	Multivariate
Glass Identification	214	10	Numerical	Real
Letter recognition	20000	16	Numerical	Integer
Fisher's Iris	150	4	Numerical	Real

AMF-IDBSCAN algorithm to DBSCAN, and the incremental DBSCAN, in terms of the Dunn Index and Davies-Bouldin criteria, as well as the change in time (milliseconds) with the increment of data in the original database, using the following equation (see Eq. 3).

$$\%\gamma changeinDB = ((newdata - olddata))/(olddata * 100) \quad (3)$$

The % delta change in the original database refers to this data increment. Experiments compare our modified incremental DBSCAN clustering algorithm to the other algorithms. The tables and figures below show how DBSCAN clustering functions with changing database data.

Remarks: The static DBSCAN algorithm was applied to 80% of each dataset, with the remaining 20% used for incrementation.

Table. II shows the results of clustering obtained using DBSCAN to group (80%) of the data for each database by changing the input parameters (MinPts and ϵ, which take the values: 8,5 ; 5,5 and 7,5 respectively) for each run:

TABLE II
THE DBSCAN ALGORITHM RESULTS

UCI Datasets	Number of *data lines*	Number of *clusters*	Evaluation	
			Davies Bould	***Dunns Index***
3D Road Network	347900	2	0.3424	0.4962
	347900	2	0.2693	0.6844
	347900	3	0.3277	0.3985
Wine	143	2	0.2948	0.3106
	143	3	0.3119	0.4534
	143	4	0.2453	0.5941
Glass Identification	173	2	0.2578	0.4343
	173	3	0.2277	0.3458
	173	4	0.2253	0.5213
Letter Recognition	16000	2	0.2213	0.4578
	16000	3	0.2158	0.6653
	16000	4	0.2343	0.3277
Fisher's Iris	120	2	0.2456	0.4856
	120	3	0.2136	0.6021
	120	4	0.2844	0.3210

We have chosen for each database a partition that gives the best values of Davies Bouldin and Dunns Index (great valuc of vD and a small value of DBI). We take the DBSCAN algorithm results (see Table. II) and apply the incremental DBSCAN algorithm [11] and our proposed modified AMF-IDBSCAN algorithm to group the 20% of new data lines for each dataset with the input parameters (MinPts and ϵ) equivalent to the line that gives the best results in vD and DBI. The results (see Table. III) are as follows:

Note that the number of clusters in the proposed modified AMF-IDBSCAN algorithm is the same or greater compared with the incremental DBSCAN algorithm of [11] results.

We opted to display the results of the first dataset to calculate the change in time (milliseconds) with the increase of data in the original database (3D Road Network). The rest is the same.

The time slowly increases with the increase of data in the original database, (as seen in Fig. 2 and Table. IV). When new data is entered into the old database, the existing and suggested incremental AMF-DBSCAN clustering algorithms are applied to the new data. These two techniques cluster new data without having to rerun the DBSCAN algorithm.

Fig. 3 and Fig. 4 show how the time in incremental databases rapidly increases as the number of data increases. By integrating the above two results (see Fig. 5), we can see that

TABLE III
COMPARISON OF THE RESULTS OF THE INCREMENTAL DBSCAN ALGORITHM [11] AND THE PROPOSED MODIFIED AMF-IDBSCAN RESULTS

UCI Datasets	**Num. of** *data lines*	**Inc DBSCAN Algorithm of [11]**		
		Num. of clusters	*Evaluation of clus*	
			Davies Bouldin	*Dunns Index*
3D Road Network	86974	2	0.2546	0.6954
Wine	35	3	0.2365	0.5958
Glass Identification	41	4	0.1986	0.6695
Letter Recognition	4000	3	0.2546	0.6579
Fisher's Iris	30	4	0.2056	0.6235
UCI Datasets	**Num.of** *data lines*	**Mod AMF-IDBSCAN**		
		Num. of clusters	*Evaluation of clus*	
			Davies Bould	*Dunns Index*
3D Road Network	86974	3	0.2179	0.7142
Wine	35	4	0.2168	0.6125
Glass Identification	41	4	0.1795	0.6845
Letter Recognition	4000	4	0.1974	0.7076
Fisher's Iris	30	4	0.1947	0.6859

TABLE IV
TIME VS. DATA IN ACTUAL DBSCAN CLUSTERING

Original data	Time (ms)
110000	32.520
120000	34.923
130000	46.739
140000	55. 063
....	....

TABLE V
TIME VS. INCREMENTAL DATA IN INCREMENTAL DBSCAN CLUSTERING [11]

Original data	Time (ms)
10000	26.762
20000	77.535
30000	173.797
40000	336.203
50000	616.283
....	....

TABLE VI
TIME VS. INCREMENTAL DATA IN MODIFIED AMF-IDBSCAN CLUSTERING

Original data	Time (ms)
10000	28.845
20000	80.632
30000	196.564
40000	388.512
50000	644.495
....	....

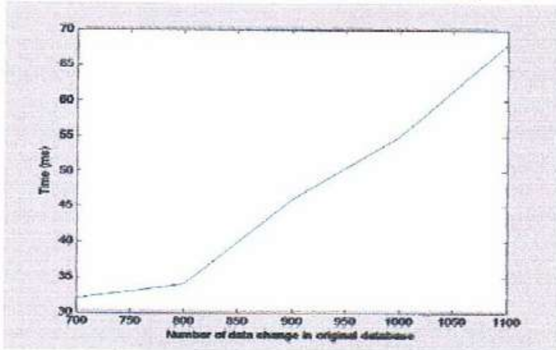

Fig. 2. Graph of actual DBSCAN result.

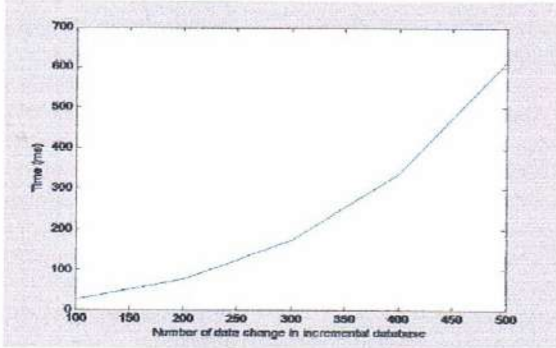

Fig. 3. Graph of existing incremental DBSCAN result.

the proposed modified AMF-IDBSCAN clustering performs better than DBSCAN clustering and incremental DBSCAN clustering of [11] for what percent of delta change [insertion of some new data items into an already existing database](see Table. VII).

γ_1 is calculated as follows: $\gamma_1 = (120000/-110000)/110000 * 100 = 9.09\%$, and we apply the same formula for the other gamma.

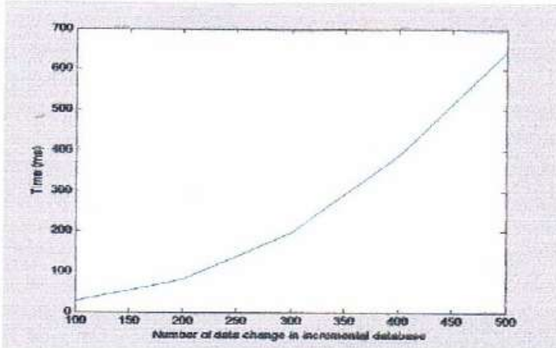

Fig. 4. Graph of proposed modified AMF-IDBSCAN result.

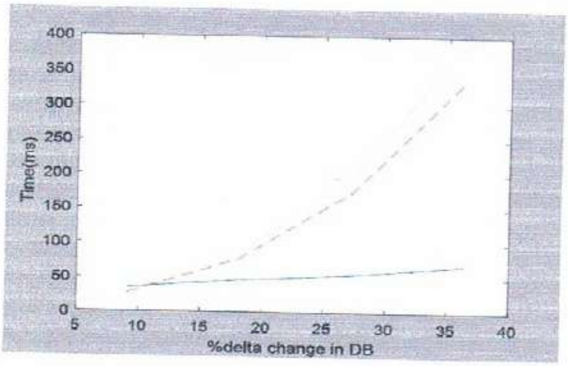

Fig. 5. Graph of actual DBSCAN vs. the two incremental DBSCAN.

TABLE VII
TIME VS. %γ CHANGE IN DB FOR THE TWO INCREMENTAL DATA DBSCAN ALGORITHMS

Actual (time ms)	% gamma Change in DB	Incr-time(ms) IDBSCAN [11]	Incr time(ms) proposed algorithm
34.923	$\gamma_1 = 9.09$	26.762	28.845
46.739	$\gamma_2 = 18.18\%$	77.535	80.632
55.063	$\gamma_3 = 27.27\%$	173.797	196.564
68.861	$\gamma_4 = 36.36\%$	336.203	388.512
....	....	....	....

C. Performance evaluation

The comparison is based on the idea that the suggested modified AMF-IDBSCAN method can be easily compared for every percent of delta change in the database. This comparison is based on the premise that each percent of delta change in the database is handled differently. The suggested developing AMF-IDBSCAN clustering algorithm is better than incremental DBSCAN clustering because it requires less time for the specific change of data in the database, whereas incremental DBSCAN [11] take significantly longer. These algorithm take longer because they need more time to correctly handle and cluster noisy data. However, the suggested modified developing AMF-IDBSCAN algorithm never wastes time dealing with outliers because it employs Adaptive Median Filtering to eliminate noise and outliers.

V. CONCLUSION AND PERSPECTIVES

In this study, we offer a modified AMF-IDBSCAN, an improved version of the original incremental DBSCAN method that incorporates the concepts of density, closeness, and noise removal. The performance of this proposed approach when combined with the Adaptive Median Filtering technique for the elimination of outliers generated by incremental DBSCAN is compared in this paper. The proposed approach outperforms DBSCAN in terms of execution time and time change (milliseconds) with data increment in the original database, according to our tests. We want to extend our research into more incremental clustering methods in the future, such as COBWEB and incremental OPTICS..... Additionally, incremental supervised algorithms such as incremental SVM, learn++, and others.

ACKNOWLEDGMENT

The authors would like to thank the DGRSDT (General Directorate of Scientific Research and Technological Development) - MESRS (Ministry of Higher Education and Scientific Research), ALGERIA, for the financial support of LISCO Laboratory.

REFERENCES

[1] Y. S. Patil, and M. B. Vaidya, "A technical survey on cluster analysis in data mining," in J. Emerg.Techn.and.Advance.Eng, vol. 2,pp. 503–513, 2012.
[2] S. ur Rehman and M. N. A. Khan, "An incremental density-based clustering technique for large datasets, " in Comput.Intel.in.Secur.for.Inf.Syst., Springer, Berlin, Heidelberg, pp. 3–11, 2010.
[3] S. Abudalfa and M. Mikki , "K-means algorithm with a novel distance measure," in.Turk.J.of.Elect.Eng.Comput.Sci, vol. 21,pp.1665–1684, 2013.
[4] K. M. Kumar and A. R. M. Reddy , "A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method, " Pattern.Recog, vol. 58, pp.39–48, 2016.
[5] P. A. Kulkarni and P. Mulay P, "Evolve systems using incremental clustering approach," Evolv.Syst, vol. 4, pp. 71–85, 2013.
[6] N. Goyal N. P. Goyal , K. Venkatramaiah, P. C. Deepak, and P. S. Sanoop , "An efficient density based incremental clustering algorithm in data warehousing environment, " in Inter.Conf.on.Comput.Eng.and Appl, IPCSIT , vol. 2, pp. 482–486, 2011.
[7] F. Tseng, D. Filev, and R. B. Chinnam , "A mutual information based online evolving clustering approach and its applications, " Evolv.Syst, vol. 8, pp. 179–191, 2017.
[8] X. Liu, Q. Yang , and L. He, "A novel DBSCAN with entropy and probability for mixed data, " Cluster.Comp, vol. 20, pp. 1313–1323, 2017.
[9] N. Suthar, P. Indr, and P. Vinit, "A Technical Survey on DBSCAN Clustering Algorithm," Int. J. Sci. Eng. Res, vol. 4, pp. 1775–1781, 2013.
[10] A. Chefrour and L. Souici-Meslati, "AMF-IDBSCAN: Incremental Density Based Clustering Algorithm Using Adaptive Median Filtering Technique," Informatica, vol. 43, 2019.
[11] S. Chakraborty and N. K. Nagwani, "Analysis and study of Incremental DBSCAN clustering algorithm," Int.J.of.Ent.Comp.and.Business.Syst., vol. 1, pp. 101–130, 2011.
[12] A. M. Bakr, N. M. Ghanem, and M. A. Ismail, "Efficient incremental density-based algorithm for clustering large datasets," J.of.Alex.Eng, vol. 54, pp. 1147–1154, 2015.
[13] P. Yada and P. Sharma, "An Efficient Incremental Density based Clustering Algorithm Fused with Noise Removal and Outlier Labelling Technique," Ind.J.of.Sci.and.Tech, vol. 9, pp. 1-7, 2016.
[14] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, " Cluster validity methods: part I," ACM.Sig.Rec, vol. 31, pp. 40–45, 2002.
[15] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, " Cluster validity checking methods: part II," ACM.Sig.Rec, vol. 31, pp. 19–27, 2002.
[16] A. L. Mary and K. S. Kumar , " A density based dynamic data clustering algorithm based on incremental dataset," 2012.