An improved DBSCAN algorithm for clustering medical applications

Mr.Prateek A. Meshram, Assistant Professor
Department of Information Technology
Rajiv Gandhi College of Engineering and Research
Nagpur, India
prateekmeshram100@gmail.com

Mrs. Pratibha Waghale, Assistant Professor
Department of Information Technology
Rajiv Gandhi College of Engineering and Research
Nagpur, India
pratibha201986@gmail.com

***Abstract*— Data clustering is proved effective to discover structures in various medical datasets. In most of the clustering algorithms, we obtain restricted clustering which means a sample should belong to a single cluster only. Most of the real world medical datasets contain the information which may belong to more than one clusters so in order to extract this information and perform clustering we use a new improved DBSCAN clustering algorithm. In this paper, a hybrid method that combines differential evolutions (D.E.) and DBSCAN. The purpose of this algorithm is to use the differential evolutions method to evaluate the clustering results and to use DBSCAN algorithm to initialize the cluster centers. The proposed method is tested by using Precision, recall and accuracy as the metric, which is observed to be the most effective metric to evaluate the clustering regardless the rag bag, homogeneity, cluster size quality, and completeness. According to results from 10 in public offered medical datasets, the planned algorithmic rule outperforms the previous algorithmic rule and may be used as an associate economical technique for clump medical datasets.**

Keyword:- Diffrential Evolution, homogeneity, clustering

I. Introduction

One of the foremost primitive human actions is grouping objects into totally different categories supported their similarity. Within the data processing domain, this task is understood as a bunch and is one among the foremost vital and helpful ideas for analyzing giant quantities of knowledge. Additional formally, a bunch are often outlined as finding heterogeneous teams of knowledge victimization some difference criterion. Bunch has been utilized in several applications from wireless detector networks[1][2]. A number of the applications of bunch algorithms in medical domain embody[3] diagnosing, biological knowledge analysis[4], medical image segmentation[5], patient database management[6] and hospital resource management[7][8].

The availability of progressively massive volume of information stirred the event of many applications in several domains like medication, biology, transportation, etc. severally of their domains, these applications share the elemental operation of information comparison[9]. Sadly, the performance of this operation is usually hindered by the dimensions, density, and non-uniformity of the information.

To overcome the said issues, many clump algorithms e.g. the k-means formula, the DBSCAN formula, and therefore the Birch formula are projected in an endeavor to classify knowledge into teams of "similar" knowledge objects whose "representative" will be accustomed accelerate the comparisons. The projected clump algorithms disagree in, however, they tackle the subsequent four requirements:

1. No a priori data regarding a number of clusters: Given the big volume of the info, a number of clusters is usually unknown prior to.

2. Discovery of clusters with discretionary shapes: this can be vital for several applications. as an example, in network observation, the distribution of connections is typically irregular. In surroundings observations, the layout of a region with similar environmental conditions may have any form[10].

3. Ability to handle outliers and noise: The presence of outliers and noise might result in misclassifications. In several applications, e.g. intrusion detection and concealment, the detection of such anomalies during a knowledge set is that the main objective of the clump method.

4. Ability to handle information density variation: Despite its multiplicity, information usually encompasses a thin house. This handicaps the characterization of clusters.

The higher than necessities are treated otherwise that cause the proposition of families of cluster algorithms; the documented square measure partitioning, gradable and density primarily based algorithms. Most of the standard cluster algorithms, as well as partitioning and gradable ones, will notice solely umbel-like clusters, however, they fail to find clusters of arbitrary shapes. to beat the issue of characterizing indiscriminately formed clusters, many algorithms use the idea of density rather than distance as a life of similarity. These algorithms outline clusters as dense regions separated by thin or low-density regions. Among the foremost well-liked density-based algorithms, one finds DBSCAN (Density primarily based spatial cluster of Applications with Noise) [11], DENCLUE (DENsity CLUstEring)[12] and OPTICS (Ordering Points to spot the cluster Structure)[13]. These density-based cluster algorithms will handle noise and turn out clusters of various

sizes and arbitrary shapes. nevertheless, they need difficulties in handling knowledge of various density and adjacent clusters. Most of them either fail to handle the variation of densities in knowledge or handle it with a time quality.

In specific, the DBSCAN formula may be a powerful spacial clump formula. However, it suffers from a high execution time, a sensibility to density parameters that has to be fastened by the user, and a coffee performance once clump information with varied density. These limits of DBSCAN are thanks to its definition of 1 international density
threshold from 2 international density parameters threshold. As a result, DBSCAN clusters solely objects settled in a very region whose density is above the pre-fixed threshold, and it classifies the remaining objects as outliers. However, clusters with totally different densities could also be rather sorted along or part classified as outliers. To classify these problems we use differential evolutions and its variants to cluster the data with varied densities especially medical data chosen in the paper. The DE-DBCSAN approach which is used in the paper is used to cluster the data with arbitrary shapes and sizes.

II. Related Work

Finding clusters with completely different sizes and shapes may be a difficult task that DBSCAN is that the most used rule. During this section, we have a tendency to 1st summary DBSCAN so we have a tendency to discuss a number of its main extensions.

2.1 DBSCAN

For a much better understanding of DBSCAN, we tend to initial review the most ideas it uses. in keeping with the DBSCAN algorithmic program, the points within the information set may be classified into 3 differing kinds [11]:

- Core Points: A data purpose p is termed a core purpose, if the quantity of points within the neighborhood of p of radius Eps exceeds some given threshold price .
- Border Points: A purpose p is named a border point, if it's variety of neighbors fewer than $Minpts$ inside Eps, however it's within the neighborhood of a core purpose.
- Noise Point: a data purpose p is termed a noise purpose, if it's neither a core purpose nor a border purpose.
- Direct Density reachability: A point o is directly density approachable from a degree p, if o is within the Eps neighborhood of p and p may be a core object.
- Density Reachability: a point p is claimed to be density approachable from some extent o with regard to Eps and , if and as long as, there exists a sequence of points $p1, p2, \dots, pn$ specified $p1 = p$, $pn = o$ and $pi + 1$ is directly density approachable from $pi\ for\ i = 1, \dots, n$.
- Density Connectivity: a point p is density-connected to some extent o with reference to Eps and $Minpts$ if there's some extent r suchboth p and o ar each density-reachable from r with reference to Eps and $Minpts$.

DBSCAN defines a cluster C with relevance 2 density parameters Eps and $Minpts$ in a very information set D as a non-empty set in D satisfying the subsequent 2 conditions:

- Maximality: if p belongs to C and o is density approachable from p with reference to Eps and , then o conjointly belongs to C.
- Connectivity: if p and o belong to C, then p is density-connected from o with relation to Eps and $Minpts$ in C .

The main steps of DBSCAN may be summarized as follows: Given an information set D, Eps and $Minpts$ as input, DBSCAN forms clusters by characteristic the Eps -neighborhood of every purpose in D. If the dimensions of the Eps -neighborhood is over , then a replacement cluster is made and DBSCAN can search all directly density approachable purposes from this core point. The method terminates once there are no new points to be intercalary to any cluster.

Overall, DBSCAN has the subsequent main advantages: it will confirm the quantity of clusters 1, it may realize clusters of arbitrary shapes and sizes 2, and it will handle noise 3. However, DBSCAN suffers from a time quality that is O for n knowledge points. associate degree economical implementation of DBSCAN supported KD-tree arrangement[14] exists with $O(n\ log\ (n))$ computing time. additionally, DBSCAN shows low performance for knowledge sets with variable densities 4 . This deficiency is as a result of the utilization of worldwide density parameters Eps and $Minpts$ whereas the density of clusters could vary. What is more, the clump results square measure sensitive to those user-fixed parameters: If these thresholds aren't adequately chosen from the beginning, DBSCAN wouldn't turn out the proper clusters. sadly, there's no clear suggests that on a way to select these thresholds. These performance drawbacks of DBSCAN prompted the proposition of many extensions among that we have a tendency to next summary the foremost standard.

III. Proposed Work

A good clustering algorithm should be able to detect the correct density levels for clustering automatically. DE-DBSCAN automatically gives us the exact values of efficient clustering with the help of differential evolutions.

3. DE-DBSCAN:

Differential Evolution (DE) Algorithm is a new evolutionary computational method for global optimization over continuous spaces proposed by Storn in 1997[15]. Differential Evolution is similar to the overall structure of the genetic algorithm [16][17]. DE consists of three basic operators: mutation, crossover, and selection. Mutation
is the most important operator in the performance of the DE algorithm because it generates new elements for the population, which may contain the optimum solution of the objective function [18][19]. In this DE-DBSCAN we have

used three different variants of DE to get the best clustering results and the best clustering outputs. The variants of the differential Evolutions which are being used in this algorithms are Rand DE, Best DE and Best of Best DE. The detailed description of all three variants of DE is shown below.

To calculate the mutation operation in the DE there are many variants, each having a different strategy to calculate the mutation operation. Some of very useful DE strategies are explained in the following[19]. Some of the strategies used in this paper are as follows.

3.1.1 Strategy/DE/Rand/1[20].
3.2.2 Strategy/DE/Best/1[21].
3.3.3 Strategy/DE/Best/2[22].

The DE-DBSCAN algorithm can be summarized as follows[23]:

Step1: Select the input parameters and initialize the population randomly.

Step2: Initialize the DE to be used for the clustering process.

Rand DE.

$$vi = xr1 + F1(xr2 - xr3) \quad (1)$$

DE Best.

$$vi = xbest + F1(xr2 - xr3) \quad (2)$$

DE Best2

$$vi = xbest + F1(xr2 - xr3 + xr4 - xr5) \quad (3)$$

Step3: Calculate the best fitness value in terms of Average of Compactness(Cavg) and keep the minimum Cave value as the Best Solution value.

Step4: Generate the mutated values with the help of mutation operation.

Step5: Perform the crossover operation and generate new trail individual values.

Step6: The generated population is used in the fitness function as the input to the DBSCAN algorithm.

Step7: The IDX generated by the DBSCAN algorithm is then passed into the Cave to find the trail individual values and that value is the minimum value of the Cave for that specific iteration.

Step 8: Choose the minimum value between the trail individual values and the corresponding individual with respect to the minimum Cave value with respect to the fitness function. Keep the best value in the best solution.

Step 9: Update the values of the Best Solutions at each iteration and also keep the best value in the Best Solution.

Step 10: Continue the above steps till the termination condition is true either Cave is zero or the max number of iterations. Otherwise go back to step2.

IV. Result Analysis

In this section the elaborated experimental results on 3 publicly obtainable medical dataset area unit provided. First, we offer the outline of the sample datasets that were retrieved from UCI repository (UCI, 2016). Next, we have a tendency to discuss the small print of the used analysis criteria followed by the elaborated results of applying the projected algorithmic rule to sample datasets.

Table 1: Description of the dataset.

Sr. No	Dataset	Number of Samples	Number of features
1	Breast Cancer Wisconsin Original	699	9
2	Heart Disease Original	303	14
3	Hepatitis	155	19

The results obtained from the PDBSCAN algorithms ate various variants of DE are shown in table 2 where Itr is represented as iterations, Eps is defined the radius of the cluster, Minpts are the minimum number of points to form a cluster. The results were calculated on the 3 variants of DE stated above which shows that the best of best DE leads us to efficient clustering with the most efficient outputs. These outputs are then compared to the prior defined algorithms such as OKM and OKM-KHM. The comparative chart of the algorithms is shown in table 3.

From the comparison given in table 3, it is clearly observed our algorithm is more efficient and gives good clustering outputs on the used medical datasets and gives better values of precision and recall than the previous working algorithms to use to implement the medical datasets. To check the efficiency of the algorithm statistical testing using Friedman test was carried out, and taking into considering the test results our algorithm gives us the efficient output. The rankings of the algorithms are as OKM-2.2143, KHM-OKM-2.2143, and our algorithm DE-DBSCAN-1.5714.

Table 2: Results of the datasets on various types of Differential Evolution (D.E)

DS	Type of D.E.	No of Itr	Eps	Min pts	Precision	Recall	Accuracy
		30	19.643	6	**0.5525**	**0.8783**	**0.8913**
Wisconsin original	**D.E Rand**	1	6.0950	10	0.3523	0.6591	0.6691
		10	4.7812	10	0.4525	0.6649	0.6745
		20	3.1354	2	0.9042	0.9856	0.9956
		30	2.7361	2	**0.9692**	**0.9952**	**0.9971**
	D.E. Best 1	1	4.4851	12	0.3500	0.6610	0.6728
		10	3.1364	3	0.4312	0.6700	0.6800
		20	3.1354	2	0.9090	0.9859	0.9971
		30	3.1453	2	**0.9529**	**0.9860**	**0.9971**
	D.E Best 2	1	5.1366	15	0.4200	0.6710	0.6801
		10	5.7311	12	0.4314	0.6800	0.6918
		20	3.1612	2	0.9100	0.9874	0.9971
		30	3.1368	2	**0.9664**	**0.9900**	**0.9981**
Hepatitis	**D.E Rand**	1	55.120	35	0.3101	0.4733	0.8017
		10	57.379	35	0.3503	0.4832	0.8324
		20	47.626	14	0.3842	0.5032	0.8679
		30	57.542	35	**0.3890**	**0.5109**	**0.8699**
	D.E. Best 1	1	58.132	33	0.3101	0.4733	0.8124
		10	55.123	32	0.3503	0.4832	0.8324
		20	61.458	36	0.3953	0.5143	0.8714
		30	41.013	35	**0.4011**	**0.5199**	**0.8801**
	D.E Best 2	1	53.817	31	0.3101	0.4738	0.8124
		10	41.034	11	0.3689	0.5032	0.8428
		20	58.700	34	0.4038	0.5279	0.8982
		30	57.542	35	**0.4109**	**0.5306**	**0.9091**
	D.E Rand	1	58.966	22	0.357	0.3776	0.6274
		10	35.530	8	0.9200	0.9664	0.9264
		20	41.534	10	0.9225	0.9664	0.9279
		30	40.756	9	**0.9249**	**0.9664**	**0.9297**

Wisconsin Diagnostic	D.E. Best 1	1	51.676	17	0.9122	0.9608	0.9174
		10	41.458	10	0.9274	0.9468	0.9244
		20	41.502	10	0.9337	0.9664	0.9315
		30	41.042	9			
	D.E Best 2	1	55.336	19	0.9149	0.9636	0.9204
		10	34.367	6	0.9244	0.9580	0.9315
		20	40.653	9	0.9274	0.9664	0.9315
		30	41.531	9	**0.9278**	**0.9668**	**0.9318**

Table 3: Comparison between Our algorithm and some previous algorithms on all the datasets with P value test.

Dataset	OKM		KHM-OKM		Our	
	Precision	Recall	Precision	Recall	Precision	Recall
Wisconsin Original	0.8471 ±0.0000	0.9846 ±0.0000	0.8471 ±0.0000	0.9846 ±0.0000	**0.9664** ±0.0001	**0.9900** ±0.0000
Hepatitis	0.6610 ±0.0020	0.8400 ±0.0074	**0.6619 ±0.0000**	**0.8661 ±0.0000**	0.4109 ±0.0002	0.5306 ±0.0035
Wisconsin Diagnostic	0.6740 ±0.0000	0.9380 ±0.0000	0.6740 ±0.0000	0.9380 ±0.0000	**0.9278** ±0.0000	**0.9668** ±0.0000

v. Conclusion

The projected DE-DBSCAN is one in every of the only and best strategies for characteristic clusters on medical knowledge. However, the OKM methodology is sensitive to the randomly chosen initial cluster centroids. Hence, during this study, we have a tendency to addressed this limitation by proposing a DE-DBSCAN rule, wherever the initial points ar chose in line with the results of bunch rule. Experimental results victimization seven in public obtainable medical datasets show that the projected methodology provides higher or comparable results compared to the initial OKM rule. moreover, we've incontestible the effectiveness of the systematic data format of OKM rule by scrutiny the target perform values at the primary iteration of the OKM rule. in line with results (except one dataset), the target perform worth of DE-DBSCAN rule was higher than original OKM rule with the random data format. although we've solely centered on the medical datasets during this study, the applications of the DE-DBSCAN methodology isn't restricted to medical domain, and it may be applied to the other domain wherever the overlapping bunch strategies ar helpful. Despite some encouraging results, there ar some limitations that require being addressed in our future work. initial of all, most of the general public datasets together with those utilized in this study ar preprocessed and so the performance of the projected methodology couldn't be absolutely understood. Experimental knowledge from real-world eventualities ought to be wont to additional analyze the performance of the projected DE-DBSCAN approach. Hence, additional studies ar needed to outline a performance metric that might expeditiously capture the performance of bunch algorithms while not wishing on the category label data. Finally, although we have a tendency to addressed one in every of the restrictions of OKM rule (sensitivity to initial cluster centers), the OKM methodology has many alternative limitations that require being addressed for a good overlapping bunch methodology. one in every of the opposite major limitations of the OKM methodology is its reliance to the geometer distance. victimization pairwise geometer distance for capturing the similarity of information points ignores the worldwide distance variation within the dataset. Some strategies like OKM- σ are projected to address this limitation NCir & Essoussi, 2013, that ought to be thought-about. An attainable future direction of this analysis is integration additional economical metaheuristic improvement rules like genetic algorithm to the projected hybrid framework rather than the iterative approach utilized in the DE-DBSCAN methodology to beat the native minimal drawback that's common among unvaried improvement strategies.

References

[1] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Comput. Commun.*, vol. 30, no. 14–15, pp. 2826–2841, 2007.

[2] Ashwini Vinayak Bhad, K. Ramteke, and I. Information Technology, R.G.C.E.R, Nagpur, "Content based image retrieval a comparative based analysis for feature extraction approach," 2015.

[3] P. G. N.S.Nithya, Dr.K.Duraiswamy, "A Survey on Clustering Techniques in Medical DiagnosisA Survey on Clustering Techniques in Medical Diagnosis A Survey on Clustering Techniques in Medical DiagnosisA," vol. 1, pp. 174–180, 2014.

[4] P. Kalyani, "Approaches to Partition Medical Data using Clustering Algorithms," *Int. J. Comput. Appl.*, vol. 49, no. 23, pp. 7–10, 2012.

[5] Z. Ma, J. M. R. S. Tavares, and R. M. Natal Jorge, "a Review on the Current Segmentation Algorithms for Medical Images," pp. 135–140, 2009.

[6] F. A. Da Veiga, "Structure discovery in medical databases: A conceptual clustering approach," *Artif. Intell. Med.*, vol. 8, no. 5, pp. 473–491, 1996.

[7] D. Dilts, J. Khamalah, and A. Plotkin, "Using cluster analysis for medical resource decision making.," vol. 15, no. 4, pp. 333–347, 1994.

[8] K. Ramteke and S. Rawat, "Lossless Image Compression LOCO-R Algorithm for 16 bit Image," *IJCA Procedding*, 2011.

[9] P. P. Pratibha Ghode, A Gaikwad, "A Keyless approach to Lossless Image Encryption," *Int. J. Adv. Res. Comput. Sci. Softw. engg*, vol. 04, no. 05, 2014.

[10] F. Cao, M. Estert, W. Qian, and A. Zhou, "Density-Based Clustering over an Evolving Data Stream with Noise," *Proc. 2006 SIAM Int. Conf. Data Min.*, pp. 328–339, 2006.

[11] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, pp. 226–231, 1996.

[12] H. Rehioui, A. Idrissi, M. Abourezq, and F. Zegrari, "DENCLUE-IM : A New Approach for Big Data Clustering," *Procedia - Procedia Comput. Sci.*, vol. 83, no. Ant, pp. 560–567, 2016.

[13] M. Ankerst, M. M. Breunig, and H. Kriegel, "OPTICS : Ordering Points To Identify the Clustering Structure," *SIGMOD '99 Proc. 1999 ACM SIGMOD Int. Conf. Manag. data*, pp. 49–60, 1999.

[14] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[15] R. Storn and K. Price, "Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces," *J. Glob. Optim.*, vol. 11, no. 4, pp. 341–359, 1997.

[16] D. Zou, J. Wu, L. Gao, and S. Li, "A modified differential evolution algorithm for unconstrained optimization problems,"

Neurocomputing, vol. 120, pp. 469–481, 2013.

[17] and F. S. Marco Locatelli, Mirko Maischberger, "No Title," *Differ. Evol. methods based local searches. Comput. Oper. Res. 43169–180, 2014.*, 2014.

[18] and Y. W. Yang Lou, Junli Li, "A Binary-Differential Evolution algorithm based on Ordering of individuals," *IEEE Nat. Comput. (ICNC), 2010 Sixth Int. Conf.*

[19] Y. A.; H. Chi, "Experimental Study on Differential Evolution Strategies," *IEEE Intell. Syst. 2009. GCIS '09. WRI Glob. Congr.*, 2009.

[20] Z. Yang, K. Tang, and X. Yao, "Self-adaptive differential evolution with neighborhood search," *2008 IEEE Congr. Evol. Comput. CEC 2008*, pp. 1110–1116, 2008.

[21] V. L. Huang, A. K. Qin, and P. N. Suganthan, "Self-adaptative {D}ifferential {E}volution {A}lgorithm for {C}onstrained {R}eal-{P}arameter {O}ptimization," *2006 IEEE Congr. Evol. Comput.*, pp. 324–331, 2006.

[22] R. Storn, "On the usage of differential evolution for function optimization," *Biennial. Conf. North Am. Fuzzy Inf. Process. Soc. 1996*, pp. 519–523, 1996.

[23] M. F. Ling Wang, Xiping Fu, Yunfei Mao, Muhammad Ilyas Menhas, "A novel modified binary differential evolution algorithm and its applications," *ELSEVIER*, vol. 98, 2012.