An Incremental Clustering Algorithm for Varied Density Data

Rimi Gupta
Sardar Vallabhbhai Patel Institute Of Technology, Vasad

ABSTRACT: *This Paper provides the methodology for generation of accurate cluster for Incremental varied density dataset. Clustering is a primary method for data mining. Data clustering can be considered as the most important unsupervised learning technique as it deals with finding a structure in a collection of unlabeled data. A Clustering is division of data into similar objects. A major difficulty in the design of data clustering algorithms is that, in majority of applications, new data are dynamically appended into an existing database and it is not feasible to perform data clustering from scratch every time new data instances get added up in the database. The development of clustering algorithms which handle the incremental updating of data points is known as an incremental clustering. It is related to the working of Incremental clustering methods and their importance on dynamic data. Here, I compared Existing Density Based Clustering Method and Incremental Density Based Clustering Method in terms of required execution time. The results are represented graphically in order to highlight the high performance of Incremental Clustering. After that applied Different Characteristics of Dataset on Incremental DBSCAN algorithm and analyze its accuracy. Analyze that Incremental Density Based Clustering cannot work with varied density datasets accurately or it cannot generate accurate clusters for varied density dataset. So, Incremental DBSCAN algorithm is applied on Incremental Varied Density Synthetic Dataset & Real Dataset and compared with Incremental Varied Density Clustering algorithm. The results prove that Incremental Varied density clustering algorithm generates accurate clusters.*

KEYWORDS: Density-based Clustering: DBSCAN; Incremental DBSCAN clustering; VDBSCAN

1. INTRODUCTION:

Data clustering is an important and extremely challenging problem. It groups data into meaningful subclasses such that the points in each subclass have high intra-class similarity and low inter-class similarity [1]. As an attractive research area in data mining, clustering is extensively used in varieties of applications such as pattern recognition, image processing, business intelligence etc.

Density based Clustering algorithm is one of the most popular algorithm for clustering data [2]. It discovers clusters of arbitrary shapes in spatial databases with noise. In incremental approach, the DBSCAN algorithm is applied to a dynamic database where data may be frequently added and the clustering discovered by DBSCAN has to be updated [3, 12]. For clustering a new appended data required to rescan newly updated datasets again and again. The Incremental DBSCAN clustering was proposed with the advantage of limited space requirement since the entire dataset is not necessary to store in the main memory and clustering newly data without rescanning a dataset so it can save a lots of time also.

Incremental DBSCAN clustering algorithm can find arbitrary shaped and differently sized clusters, handle noisy data also. In the meanwhile, due to adoption of global parameters, it fails to identify clusters with varied densities unless the clusters are clearly separated by sparse regions.

In this paper, proposed a new Incremental clustering algorithm to discover clusters of different densities. The experimental results show the superiority of this algorithm.

The rest of the paper is organized as follows: Section 2 reviews Incremental DBSCAN clustering algorithm and its analysis with different characteristics of datasets. Section 3 represents details of proposed algorithm. Section 4 shows a experimental results. At the end,

2. INCREMENTAL DBSCAN CLUSTERING ALGORITHM AND ANALYSIS

- **Introduction to Incremental DBSCAN Clustering Algorithm:**

The term incremental means "% of change in the original database" i.e. insertion of some new data items into the already existing clusters. Such as, [3]

%Change in DB = (New Data – Old Data) / (Old data) * 100

Incremental clustering is the process of updating an existing set of clusters incrementally rather than mining them from the scratch on each database update.

In Incremental approach, DBSCAN algorithm is applied to a dynamic dataset where data may be updated frequently. In this approach, first compute the means between every core objects of clusters and the new coming data and insert the new data into a particular cluster based on the minimum mean distance. The new data which are not inserted into any clusters, then treated as a noise or outliers [3].

- **Incremental DBSCAN clustering algorithm steps [3].**

1) Use Original DBSCAN clustering algorithm on dataset and get clusters.
2) Find the mean of each cluster.
3) Find distance between new data item and mean of clusters.
4) Find minimum distance from each cluster and also check whether this distance is less Than or equal to given epsilon.
5) If condition is satisfied then add new data item in appropriate cluster.
6) Otherwise that point treat as an outlier.
7) Update the existing clusters

- **Performance Evolution of Incremental DBSCAN Clustering**

In this paper I have compared DBSCAN and Incremental DBSCAN with respect to time analysis. For analysis purpose I have use **German Credit Card** Dataset in Incremental way. The dataset taken from UCI Repository dataset [18].

Dataset Name: **German Credit Card Data**

No. of Instances: **1000**

No. of Attributes: **25**

First apply DBSCAN clustering algorithm on 500 data value. Then take its result as an input to the Incremental DBSCAN clustering algorithm and take incrementally new data like, 50,100,150,200,250… and analyze the time for both algorithms

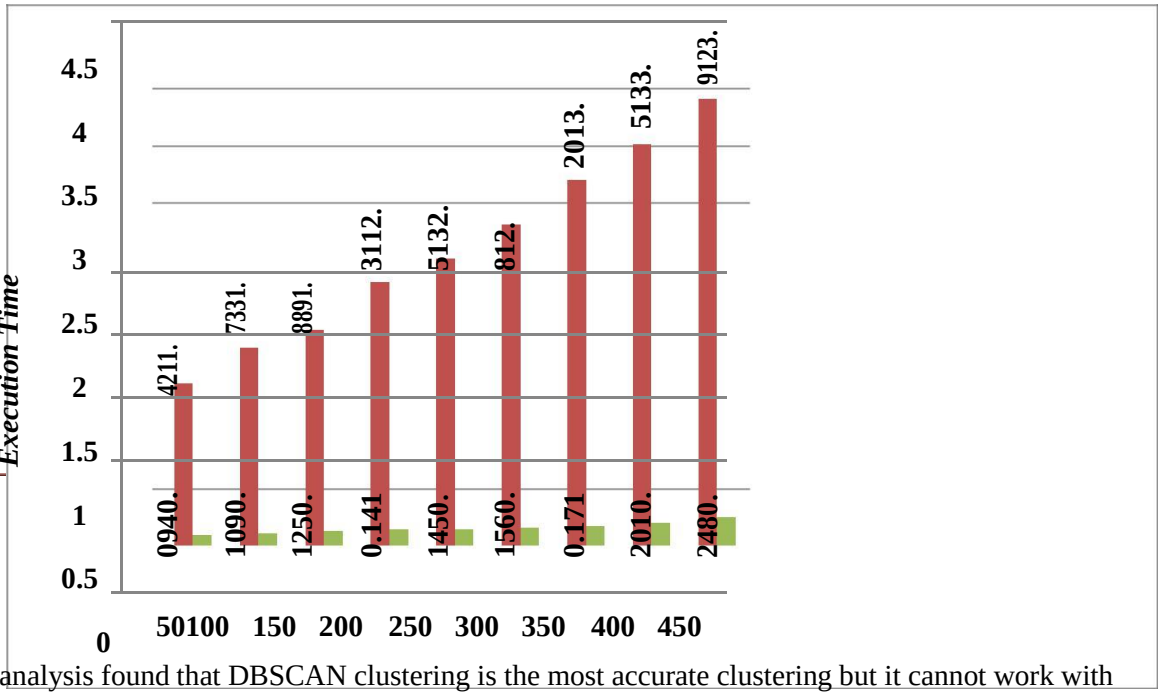
- **Time for first 500 data value in DBSCAN: 3.56 ms**

Table 1: Time Analysis of Incremental German Credit Data

New Data Item	No. of data run on DBSCAN	Execution Time (ms)	No. of data run on Incremental DBSCAN	Execution Time (ms)
50	550	1.421	50	0.094
100	600	1.733	100	0.109
150	650	1.889	150	0.125
200	700	2.311	200	0.141

250	750	2.513	250	0.145
300	800	2.81	300	0.156
350	850	3.201	350	0.171
400	900	3.513	400	0.201
450	950	3.912	450	0.248

Based on Table1, the results are displayed in the graphical format below. Graph indicates that Incremental DBSCAN is better than DBSCAN algorithm



From above analysis found that DBSCAN clustering is the most accurate clustering but it cannot work with Incremental data with efficient time where Incremental DBSCAN clustering can work very efficiently where data are added into exiting clusters and clusters are updated.

- **Performance Evaluation of Incremental DBSCAN Clustering Algorithm on Different Characteristics of Datasets**

Incremental DBSCAN clustering should be generate accurate clusters with different characteristics of datasets like generate arbitrary shapes clusters, Handle noisy data very efficiently, handle high dimensional dataset and varied density dataset. I have applied
An Incremental Clustering Algorithm for Varied Density Data
Incremental DBSCAN Clustering algorithm on all the types of datasets and check accuracy whether it can generate accurate clusters or not.

- **Generate Arbitrary Shapes Clusters**

Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. Consider sensors, for example, which are often deployed for environment surveillance. Cluster analysis on sensor readings can detect interesting phenomena. We may want to use clustering to find the frontier of a running forest fire, which is often not spherical. It is important to develop algorithms that can detect clusters of arbitrary shape.

- **Handle Noisy Data**

Most real-world data sets contain outliers and/or missing, unknown, or erroneous data. Sensor readings, for example, are often noisy, some readings may be inaccurate due to the sensing mechanisms, and some readings may be erroneous due to interferences from surrounding transient objects. Clustering algorithms can be sensitive to such noise and may produce poor-quality clusters. Therefore, we need clustering methods that are robust to noise.

- **Handle High Dimensional Data**

A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only two to three dimensions. Finding clusters of data objects in high- dimensional space is challenging, especially considering that such data can be sparse and highly skewed.

- **Varied Density Data**

A database may contain different density level. All the dataset are distributed uneven manner. There will be some distinct variation, depending on the densities of the clusters; the range of variation may not be huge while a sharp change is expected to see between An Incremental Clustering Algorithm for Varied Density Data two density levels. Like an example Population of literate and illiterate girls and boys is varied data. Clustering algorithm should be clusters this data accurately according to its varied density level.

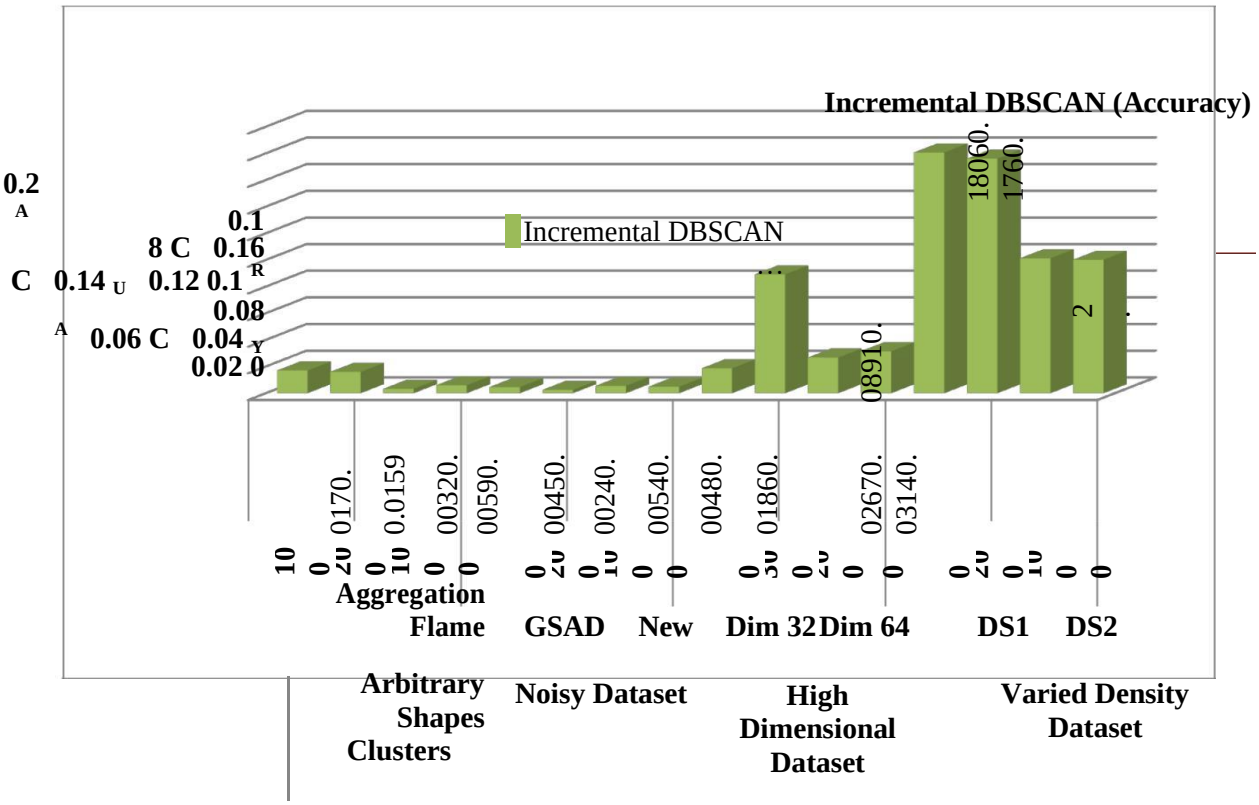
Table 2: Different Datasets Information (UCI Repository Dataset)

Sr. No	Characteristics of Dataset	Dataset Name	No. of Attributes	No. of Instances
1	Arbitrary Shapes Clusters	Aggregation	2	800
		Flame	2	800
2	Noisy Data	Gsad	7	1000
		Newer	9	800
3	High Dimensional	Dim 32	32	1024
		Dim 64	64	1024
4	Varied Density Data	DS1	2	800
		DS2	2	1000

Table 3: Experiment Analysis of Incremental DBSCAN clustering Algorithm

Characteristics of Dataset	Data Set Name	No. of Data	Incremental DBSCAN (Accuracy)
Arbitrary Shapes Clusters	Aggregation	100	0.0170
		200	0.0159
	Flame	100	0.0032
		200	0.0059
Noisy Dataset	GSAD	100	0.0045
		200	0.0024
High Dimensional Dataset	Nwe	100	0.0054
		200	0.0048
	Dim 32	200	0.0186
		300	0.0891
	Dim 64	200	0.0267
		300	0.0314
Varied Density Dataset	DS1	100	0.1806
		200	0.1760
	DS2	100	0.1012
		200	0.1001

Incremental DBSCAN Clustering algorithm is most powerful technique for updating clusters in case of new data items are added into the datasets. It does not require any prior knowledge for clustering process.
By applying the Incremental DBSCAN Clustering algorithm on different characteristics of dataset I had found that it can generate Arbitrary Shaped Clusters, Handle Noisy Dataset, Handle High Dimensional Dataset very efficiently but It cannot generate accurate clusters when varied density of datasets are applied.
Based on Table 3, the Results are displayed in the graphical format below. Graph indicates that Incremental DBSCAN can‟t generate accurate clusters on varied density data



3. PROPOSED ALGORITHM:

- **An Incremental Clustering Algorithm for Varied Density Data**

In this paper, introduce Varied Density Incremental Clustering algorithm, which is a new improved algorithm of Incremental DBSCAN algorithm that has limitation of single global parameter for generate a clusters. In this algorithm used different epsilon value for each density level set.

Input:

1. Set K clusters from VDBSCAN algorithm.
2. Set all new data objects as unclassified, then [Process Data Set]

[Process Data Set]

3. For each data object X in the data set

If X is unclassified then [Expand Cluster (K, X)]

[Expand Cluster(K, X)]

Set Counter=0

4. M=Find Mean of each K clusters.
5. Dist (M, X) = Find distance between Mean and new data item.
6. Distmin (M, X) = Find Minimum distance among all clusters.
7. Arrange all Epsilon values in ascending order
8. For each epsilon (I =1 to n)

9. If ((Distmin (M, X) <= epsilon)

Then add Point X in corresponding cluster.

Counter =1; Break;

10. End for Loop

11. If (counter=0) add Point X as an outlier.

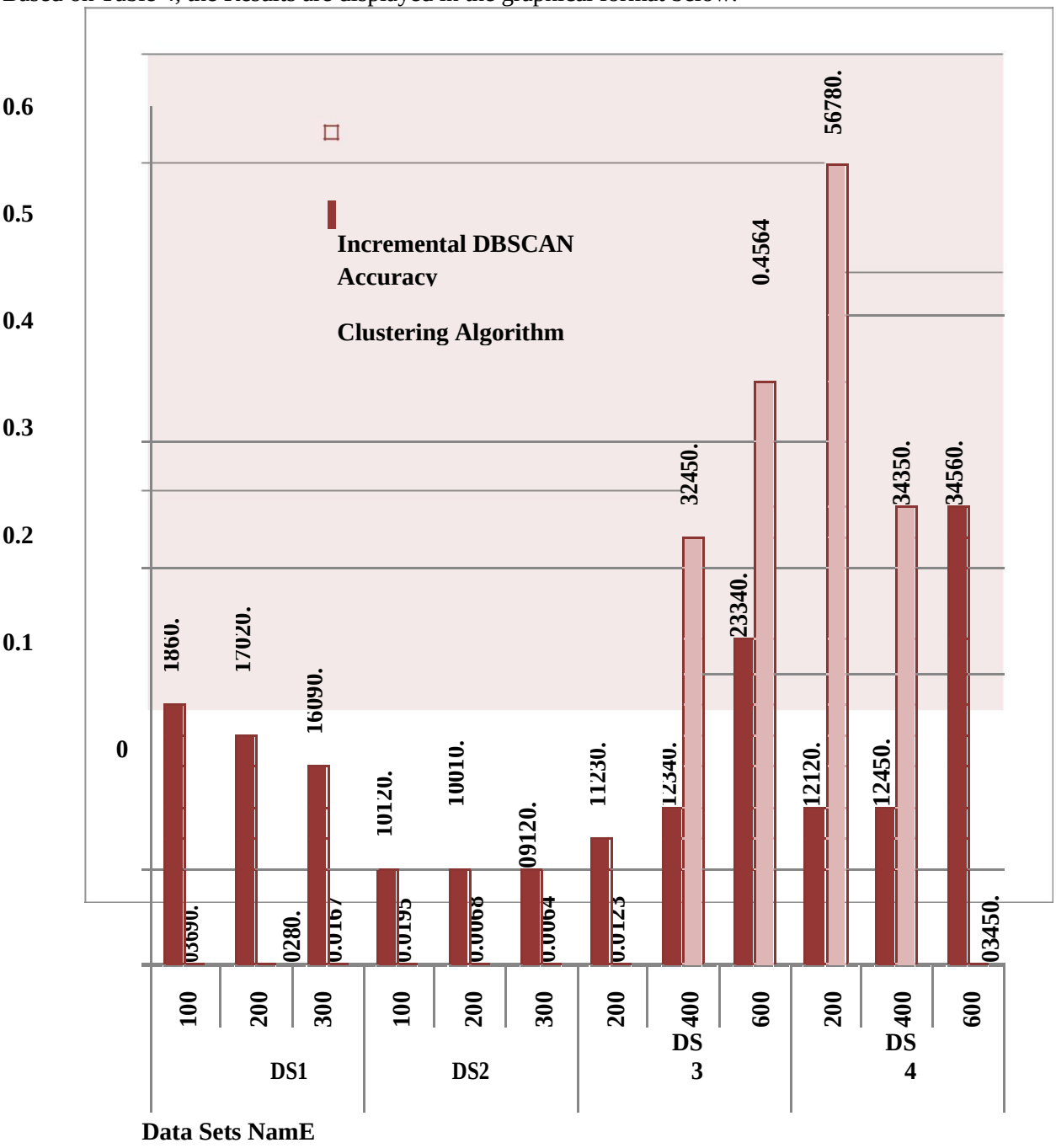
Experiments and Analysis

To evaluate the performance of Incremental Varied Density Clustering Algorithm on Varied Density Synthetic two Dimensional Datasets (DS1, DS2) with 800 instances and 8 Dimensional (DS3), 12 Dimensional Dataset(DS4) with 1000 instances [18,19] compare the results with Incremental DBSCAN clustering algorithm. Here finding an accuracy of generated clusters in both the algorithm. For the accuracy measurement used Internal Evaluation (Davies- Bouldin Index) which value is less than zero or smallest then considered the accurate clusters. So, using Davies – Bouldin Index validation compares the accuracy of both the algorithm on varied density dataset.

Table 4 Accuracy comparison between Incremental DBSCAN and Incremental Varied Density

Data Set Name	No. of Data	Incremental DBSCAN Clustering Algorithm Accuracy	Incremental Varied Density Clustering Algorithm Accuracy
DS1	100	0.186	0.0369
	200	0.1702	0.028
	300	0.1609	0.0167
DS2	100	0.1012	0.0195
	200	0.1001	0.0068
	300	0.0912	0.0064
DS3	200	0.1123	0.0123
	400	0.1234	0.3245
	600	0.2334	0.4564
DS4	200	0.1212	0.5678
	400	0.1245	0.3435
	600	0.3456	0.0345

Based on Table 4, the Results are displayed in the graphical format below.



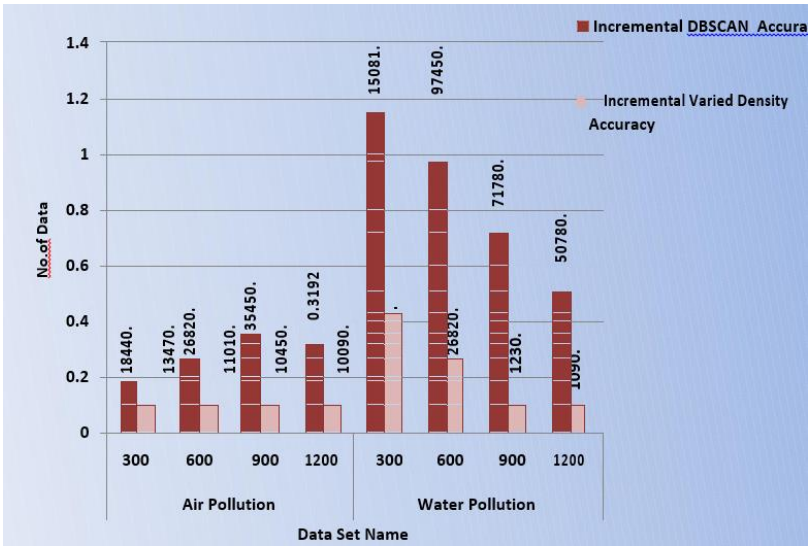
4. PERFORMANCE ON REAL-WORLD DATASETS

In order to confirm the conclusions, evaluate the performance of proposed algorithm and Incremental DBSCAN on real-world datasets called Air Pollution and Water Pollution both taken from www.cpcb.nic.in [20] Davies bouldin Index is used to compare the clustering outputs of the two methods.

Table 5 Accuracy comparison between Incremental DBSCAN and Incremental Varied Density on Real World Datasets

Air Data Set Name	No. of Data	Incremental DBSCAN Accuracy	Incremental Varied Density Accuracy
Air Pollution	300	0.1844	0.1347
	600	0.2682	0.1101
	900	0.3545	0.1045
	1200	0.3192	0.1009
Water Pollution	300	1.1508	0.4312
	600	0.9745	0.2682
	900	0.7178	0.123
	1200	0.5078	0.109

Based on Table 5, the Results are displayed in the graphical format below.



5. CONCLUSION AND FUTURE WORK:

Existing Incremental DBSCAN clustering algorithm is better than DBSCAN clustering algorithm to clusters newly added data efficiently. Apply Incremental DBSCAN clustering algorithm on Different Characteristics of Datasets. In that conclude that Incremental DBSCAN clustering can work efficiently and generate accurate clusters on Different Shapes clusters, Noisy Dataset, High Dimensional Dataset. But it cannot generate accurate

clusters in case of varied density dataset. So, by applying Incremental Varied Density clustering algorithm can work efficiently and generate accurate clusters than Incremental DBSCAN clustering algorithm on varied density dataset. Excellent performance on Synthetic 2-D, 8-D, 12-D datasets and Real world (Air pollution and Water Pollution) datasets confirms its effectiveness.

In future work, I am planning to apply Incremental Varied Density Clustering algorithm on Categorical type and mixed type datasets. And Proved that algorithm can generate accurate clusters for any type of varied density datasets.

REFERENCES

1) F. Knoll "Survey of Clustering Data Mining Techniques" Pavel Berkhin Accrue Software, Inc. Pavel Berkhin , Accrue Software, 1045., San Jose, CA, 95129;
2) Manish Varma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta, " A Comparative study of various clustering algorithms in data mining", International Journal of Engineering Research and Applications, 2012
3) Prof. Sanjay Chakra borty and Prof. N.K. Nagwani "Analysis and study of Incremental DBSCAN Clustering Algorithm ", International Journal of Enterprise Computing and Business
4) Prof. Sanjay Chakra borty and Prof. N.K. Nagwani "Analysis and study of Incremental K-means Clustering Algorithm ", International Journal of Enterprise Computing and Business
5) Martin Ester, Hans-Peter Kriegel, Jorge Sander, Michael Wilmer, Xiaowei Xu " Incremental Clustering for Mining in a Data Warehousing Environment", Proceedings of the 24th VLDB Conference, New York, USA, 1998.
6) Prof. Sanjay Chakra borty and Prof. N.K. Nagwani and Prof. Lopamudra Dey " Performance Comparison of Incremental K-means and Incremental DBSCAN algorithms ", International Journal of Computer Application, 2011.
7) Martin Ester, Hans-Peter Kriegel, Jorge Sander, Xiaowei Xu " A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).
8) M.Parimala, Daphne Lopez, N.C. Senthil Kumar "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases ", International Journal of Advanced Science and Technology, 2011.
9) Peng Liu, Dong Zhou, Naijun Wu "Varied Density Based Spatial Clustering of Application with Noise", IEEE, 2007
10) Chen Xiaoyun, Min Yufang, Zhao Yan,Wang Ping "GMDBSCAN: Multi Density DBSCAN Cluster Based on Grid".
11) Angel Lathat Maryn, K.R. Shankar Kumar, "Evaluation of Clustering Algorithm with Cluster Validation Metrics", European Journal of Scientific Research, 2012.
12) L. Kaufman and P. Rousseeuw, "Finding groups in Data: an Introduction to cluster," John Wiley & Sons, 1990.

BOOKS

13) G.K Gupta, "Introduciton to Data Mining with Case Studies", PHI (2006)
14) J.Han and M.Kamber "Data Mining Concepts and Techniques".
15) Ian H. Witten & Eibe Frank "Data Mining Practical Machin Learning Tools and Techniques
16) Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl, "Cluster analysis" 5th Edition
17) Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining," Pearson Addison Wesley, 2006.

WEB SITES

18) http://archive.ics.uci.edu/ml/datasets.html
19) http://cs.joensuu.fi/sipu/datasets/
20) http://en.wikipedia.org/wiki/Cluster_analysis
21) http://cpcb.nic.in/