IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

A CRITICAL REVIEW ON DENSITY-BASED CLUSTERING ALGORITHMS AND THEIR PERFORMANCE IN DATA MINING

Jayasree Ravi[1], Sushil Kulkarni[2]
[1]Department of Computer Science, University of Mumbai, India
[2]Department of Computer Science, University of Mumbai, India

Abstract

Mining of Spatial databases has been a subject of interest and a topic of research in recent times. Social Media are the vast sources of geo-tagged information. This massive database needs to be mined in an effective way to arrive at interesting patterns and for detecting events. Expected effectiveness is achieved only when appropriate techniques are applied to different kinds of data available over the social media platforms. This paper aims at giving insights about various density-based clustering algorithms, their domain-specific applications, datasets used, methods of data extraction. This paper also focuses on the performance evaluation of the algorithms which are part of this survey.

Index Terms: Clustering algorithms; Applications; Mining massive data; performance evaluation of algorithms;

1. INTRODUCTION

Emerging technologies and their associated devices have led to the generation of massive data over the past few years. Because of this, there is a dire need to have more progressive and efficient predictive intelligent models to meet up to the demands that will arise in future. (Al-Jarrah et al., 2015) With the recent growth of social media platforms, there has been vast amounts of data which can be mined to study various aspects such as sentiments, behavior, patterns and influence. For the effective study, we need to implement appropriate algorithm depending on the types of data. Through this data, we can discover knowledge based on the available data. For instance, just with user profile, we can find interesting patterns on the demographical details of the user. Using the textual information, we can find sentiment analysis which again is a very useful knowledge for marketing and behavior analysis.

This type of information is in any form – images, videos, text, geocodes and so on. For extracting patterns from such information, we need to use unsupervised algorithms. Clustering is one of the most popular unsupervised algorithms.(Leskovec et al., 2014)

The rest of the paper is organized as follows. Section 2 briefly discusses different types of clustering algorithms and highlights DBSCAN(Martin Ester, Hans-Peter Kriegel, Jiirg Sander, 1996). Section 3 deals with review of various research paper related to clustering algorithms and especially DBSCAN, its variations and improvisations on handling varied data sets, handling high dimensional data. Section 4 briefs the motivation behind the choice of research work that has been considered in this paper. Section 5 covers the comparison of DBSCAN with other improved models. Section 6 gives a brief idea about different aspects of the algorithms considered for this work. Finally, section 7 concludes the work with future scope of study.

2. CLUSTERING ALGORITHMS

Clustering Algorithms can be broadly classified as follows. Partition-Based, Hierarchical, Grid-Based and Density-Based(Jiawei et al., 2012).

2.1 PARTITION-BASED METHODS

K-Means clustering is a classic example of Partition-based clustering algorithm. In this clustering method, the number of clusters is got as input and the dataset is grouped under one of the clusters. There has been various improvements on this clustering technique. One such model is mentioned in (Pambudi et al., 2021) where the authors have applied grey-wolf optimizer for segmentation of brain MRI. Here, the number of clusters have to be pre-determined and this algorithm is not sensitive to noise. These two challenges of K-Means are overcome be Density-based clustering. The study handled in (Murugan & Rathna, 2019) discusses about the impact of fuzzy privacy preservation method in clustering. Sentiment analysis using K-Means clustering is discussed in (Harshini & Gobi, 2020).

2.2 HIERARCHICAL-BASED METHODS

Hierarchical clustering method demonstrates the similarity of clusters as a tree that is called dendrogram. The clusters nested in the dendrogram represent the related clusters belonging to a dataset. There are two types of algorithms in the hierarchical method: 1. Agglomerative method 2. Divisive method. Dendrogram can establish both methods. Hierarchical clustering approach uses different limitation to choose locally which cluster should be combined at each step. Agglomerative method takes into account each point as cluster, and it merges the points until final cluster is created. BIRCH (Zhang et al., 1997) is one of the main extensions of agglomerative algorithm. In divisive algorithm, all the data points are assumed as part of only one cluster. Later, the data points are subdivided into new smaller clusters until the final desired result is obtained.

2.3 GRID-BASED METHODS

It splits the data points into predefined number of cubicles that forms an arrangement of grids. The main advantage of this method is that the speed of processing is high as it only depends on numbers of cells and not dependent on number of objects. It is an effective method to many spatial data mining problems. CLIQUE(Agrawal et al., 1998) algorithm is an example of this method.

2.4 DENSITY-BASED METHODS

Density-based spatial clustering of applications with noise (DBSCAN)(Martin Ester, Hans-Peter Kriegel, Jiirg Sander, 1996) and its versions are more effective in detecting clusters with arbitrary shapes along with handling of noisy data where the prior knowledge of the number of clusters is also not needed to be specified. As density-based clustering can be used for variety of applications and is still a popular clustering algorithm, this study focuses on this clustering algorithm.

3. RELATED WORKS

There have been many review papers on Clustering Algorithms in general and Density-based clustering in particular. The details about different review papers are tabulated below in Table.1.

Table.1. Different Review Works on Clustering Algorithms

Sl. No.	Paper	Year	Clustering Algorithms Reviewed	Authors
1.	Critical Analysis of DBSCAN Variations(Ali et al., 2010)	2010	Density-based Clustering Algorithms	T. Ali, S. Asghar, and N. A. Sajid
2.	A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases(Parimala et al., 2011)	2011	DBSCAN, VDBSCAN, DVBSCAN(Ram et al., 2010) ST-DBSCAN(Birant & Kut, 2007), DBCLASD(Xu et al., 1998)	M. Parimala, D. Lopez, and N. C. Senthilkumar
3.	Comparison of Different Clustering Algorithms using WEKA Tool(Kakkar & Parashar, 2014)	2014	K-Means, EM, DBSCAN	P. Kakkar and A. Parashar
4.	A survey on density-based clustering algorithms (Loh & Park, 2014)	2014	Density Based Clustering Algorithms	W. K. Loh and Y. H. Park
5.	A Brief Survey on Clustering Algorithms in Data Mining(Kankal et al., 2017)	2017	Hierarchical-based, Partition-based, Density-based, Grid-based clustering methods	S. S. Kankal, A. R. Dhakne, and Y. R. Tayade
6.	A Comparative Quantitative Analysis of Contemporary Big Data Clustering Algorithms for Market Segmentation in Hospitality Industry (Bose et al., 2017)	2017	DBSCAN, OPTICS and Variants of DBSCAN	A. Bose, A. Munir, and N. Shabani
7.	Data clustering algorithms : A second look(Alraba & Al-refai, 2018)	2018	Hierarchical-based, Partition-based, Density-based, Grid-based clustering methods	Y. Alraba and M. Al-refai
8.	Study of Clustering Methods in Data Mining(Anitajesi & Arumaiselvam;, 2018)	2018	Hierarchical-based, Partition-based, Density-based, Grid-based clustering methods	Anitajesi and Arumaiselvam
9.	Review on Density Based Clustering Algorithms for Big Data(Lakshmi et al., 2018)	2018	DBSCAN, DENCLUE(Hinneburg & Keim, 1998), OPTICS(Ankerst et al., 1999)	M. Lakshmi, J. Sahana, and P. Venkatesan
10.	Spatiotemporal data clustering: A survey of methods(Z. Shi & Pun-Cheng, 2019)	2019	hypothesis-based clustering method and partition-based clustering method	Z. Shi and L. S. C. Pun-Cheng

11.	Performance evaluation and comparison of clustering algorithms used in educational data mining(Valarmathy & Krishnaveni, 2019)	2019	EM, CLOPE, CLARA COBWEB, Filtered Cluster, Farthest First, K-Means, DBSCAN	N. Valarmathy, S. Krishnaveni
12.	Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses(Ahmed et al., 2020)	2020	K-Means, DBSCAN OPTICS	M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin
13.	A survey of density based clustering algorithms(Bhattacharjee & Mitra, 2021)	2021	Density Based Clustering Algorithms	P. Bhattacharjee and P. Mitra

Although these surveys talk about various clustering algorithms, there has not been any study that is based specifically on the latest improvisation of density-based clustering. This paper analyses the working of seven such studies which are built on the existing density-based clustering algorithm.

Researchers have come up with various techniques in mining of such huge data. One such algorithm is Density-based clustering. The pioneer in density-based clustering is DBSCAN(Martin Ester, Hans-Peter Kriegel, Jiirg Sander, 1996).

3.1 REVIEW ON DBSCAN

This work not merely discusses working and results of existing models related to density-based clustering, but also does a thorough critical analysis of the same to learn their effectiveness and efficacy on various datasets.

The key idea of density-based clustering is that for each data point of a cluster the neighbourhood of a given radius $\varepsilon > 0$ has at least a minimum number of data points i.e., the cardinality of the neighbourhood has to exceed a given threshold.

3.2 NOTATIONS

The commonly used notations are listed in Table.2.

Table.2. Notations used

Notation	Description
D	Set of Data points
p, q	Data points ∈ D
E	E ⊂ D
$y_1,y_2,\dots y_n$	Data points ∈ E
ε	Radius from a point
MinPts	Minimum number of points within the ε radius
n	|D| Number of Data points
m	|E| Number of Data points in the subset of dataset

3.3 DEFINITION

The following Definitions explain the idea of density-based clustering (Martin Ester, Hans-Peter Kriegel, Jiirg Sander, 1996)

3.3.1 – ε -neighbourhood: The neighbourhood of a data point p ∈ D within a radius $\varepsilon > 0$ is referred to as the ε-neighbourhood of the point

3.3.2 – Core point: If the ε -neighbourhood as defined above of a data point p has at least *MinPts* number of data points, then p qualifies to be a core point

3.3.3 – Border point: If a data point q belongs to the ε neighborhood of a core point p but has a smaller number of points than *MinPts* within its own radius ε, then it is referred to as border point

3.3.4 – Directly density-reachable: A data point q is directly density-reachable from a data point p, if q falls within the ε -neighbourhood of p, and also p is a core point

3.3.5 – Density-reachable: If there is a collection of data points Let p and q belongs to D. Let E={$y_{1,\dots}y_n$ } subset of D between p and q where , such that y_1 is directly density reachable to p, y_2 is directly density reachable to y_1, and so on, and y_n is directly density reachable to q, then q is density-reachable to p with respect to ε and *MinPts*

3.3.6 – Density-connected: A data point p is density-connected to a data point q with regards to ε and *MinPts*, if there is an object y in D such that both p and q are density-reachable from y with respect to ε and *MinPts*.

This algorithm classifies every point either as a core point, border point or noise. Clusters are formed based on the density.

The advantages of DBSCAN are that the clusters can be of arbitrary shapes. Algorithm deals with noise. Prior knowledge of number of clusters need not be known. The disadvantages of DBSCAN are that the parameters needed for clustering need to be decided to ensure proper clustering, its worst-case time complexity is $O(n^2)$, n being the number of data points. This is not suitable for high-dimensional data and it is insensitive towards varied density data.

3.4 EVOLUTION OF DBSCAN ALGORITHM

GDBSCAN (Sander et al., n.d.) generalized DBSCAN by extending the concept of neighborhood over the traditional *E* -neighborhood and by using different measures to define the ‘‘cardinality’’ of the neighborhood. The novel method proposed in (Viswanath & Pinkesh, 2006) uses two types of prototypes, one for reducing the time requirement, and the other for reducing the deviation of the result. Prototypes are arrived using leaders clustering method. This derives a two-level hierarchy with different values for threshold. This paper presented a scalable hybrid clustering method to get density based clusters of arbitrary shape. As a first step, two types of prototypes are derived using leaders clustering method. The prototypes are utilized by the traditional DBSCAN. The proposed l-DBSCAN method is proved to be faster than DBSCAN using the entire data set. The authors of (Viswanath & Pinkesh, 2006) have used fast clustering to derive prototypes called leaders in (Viswanath & Suresh Babu, 2009). The authors have improved their own version of previous papers in the following ways – Theoretically establishing the properties of leaders, establishing the working of proposed method, Analysis of the proposed model using rough set theory. They have also established that rough DBSCAN running time is linear as compared to DBSCAN which is quadratic. The paper (Peng et al., 2007) deals with datasets of varied densities and thus makes the density-based clustering deal with data of varied densities. The algorithm proposed in (Tran et al., 2013) revises the concept of DBSCAN and modify the algorithm in order to achieve a better performance. This paper proposes a solution which works on spatial and non-spatial data types and addresses the issue of border data points belonging to adjacent clusters. It uses core-density-reachable chains which involves only core points instead of the traditional density-reachable objects. The border object is allotted to a cluster to which the core-density-reachable chain it belongs to. This algorithm retains the key concepts of the DBSCAN algorithm, with an additional potential to improve the results of clustering by solving the issue of border objects. The study mentioned in (Naik Gaonkar & Sawant, 2013) will automatically select the input parameters which identifies density varied clusters The algorithm discussed in (Soni & Ganatra, 2016a) presents incremental DBSCAN which can be used for multiple data objects at the same time, named MOiD (Multiple Objects incremental DBSCAN). The paper (Soni & Ganatra, 2016b) proposed a mathematical way of calculating one of the input parameters of Density-based clustering – *ε* value through K-nearest neighbor. The algorithm devised by the authors of (Priyadarshini et al., 2016) calculates both the parameters of density-based clustering. In the model proposed by authors of (Ozkok & Celik, 2017), there is a methodology AE-DBSCAN to automatically estimate the value of neighborhood radius. The model uses the k-distgraph to find out the slopes present in the density of the dataset, find the mean and standard deviation of all non-zero slopes, find the first slope which is above mean and standard deviation and assign it as *ε* value. There is a research work (Nguyen & Shin, 2017), which points out the shortcomings of DBSCAN and its variations. It discovers that while the region around a Point of Interest generally includes points that contain and do not contain annotated Point of Interest terms (denoted as Point-relevant and Point-irrelevant points, respectively), DBSCAN takes only into account only the point-relevant in the clustering process. To solve this problem, they have introduced a new algorithm, named DBSTexC, for density-based clustering on Twitter, which incorporates text information into the DBSCAN model to avoid geographical regions with numerous irrelevant geotagged posts. HDBSCAN (McInnes & Healy, 2017) was proposed by generating a hierarchy in density-based clustering and then filtering clusters based on stability measures. Considering only one type of input data (e.g., the tweets relevant to a POI) may lead to formation of inaccurate clusters due to noisy data. The model DBSCAN++ proposed in (Jang & Jiang, 2019) addresses the issue of run-time complexity of DBSCAN. Event detection using Density-based clustering is done in (Huang et al., 2018) and (Ghaemi & Farnaghi, 2019). Density-based clustering is applied on social network data for various applications as mentioned in (J. Shi et al., 2014) and (Vu et al., 2016)

4. MOTIVATION

From the literature review, we understand that there has been no studies or survey on the improved versions of Density-based clustering algorithms. Of all the improvised versions of DBSCAN, we have identified 7 algorithms, which addresses the shortfalls of DBSCAN. The reasons for choosing the algorithms are given below.

VDBSCAN - Varied Density Based Spatial Clustering of Applications with Noise(Peng et al., 2007) model finds out the shortcomings of existing density-based algorithms while finding out all the meaningful clusters for datasets with different densities. With multiple values of *ε*, it is easy to find out clusters with varied densities at the same time. AGED (Automatic Generation of *ε* for DBSCAN)(Soni & Ganatra, 2016b) handles two of the challenges faced by traditional density-based algorithms – Handling Different Densities present in datasets and Determining the input parameters. This paper points out that the density-based algorithm is sensitive to even a small variation of the input parameter values. So, this paper proposed a mathematical way of calculating one of the input parameters of Density-based clustering – ε value through K-nearest neighbor. This paper uses min-max normalization followed by binning to arrive at different ε values for different densities. Adaptive DBSCAN proposed in (Priyadarshini et al., 2016) also finds a way to calculate the ε value and the corresponding *MinPts* value. AE_DBSCAN algorithm proposed by authors of (Ozkok & Celik, 2017) also estimates the ε value by proposing 3 strategies. Density-Based Spatial–Textual Clustering on Twitter (DBSTexC) (Nguyen & Shin, 2017) discovers that while the region around a Point of Interest generally includes geo-tags that contain and do not contain annotated Point of Interest keywords (denoted as POI-relevant and POI-irrelevant geo-tags, respectively), DBSCAN takes only into account the former in the clustering process. This algorithm takes into account not only relevant points of interest but also puts a threshold limit on irrelevant information that can be present in the *ε* neighborhood. This results in better clustering. DBSCAN++: Towards fast and scalable density clustering(Jang & Jiang, 2019) is based on the notion that There need to be only a subset of data points for which the density estimates need to be computed. The authors have proposed two strategies to choose these points - uniform and greedy K-center-based sampling. Through this algorithm, the number of data points to be considered in calculation is reduced, which in turn reduces the time of execution. In the model VDCT proposed in (Ghaemi & Farnaghi, 2019), the authors have taken heterogeneous twitter data for clustering. The authors have taken into account both spatial information and textual information for clustering.

This paper compares the working of traditional DBSCAN with its seven improvised versions – VDBSCAN (Peng et al., 2007), AGED(Soni & Ganatra, 2016b), Adaptive DBSCAN (Priyadarshini et al., 2016), AE-DBSCAN(Ozkok & Celik, 2017), DBSTexC(Nguyen & Shin, 2017), DBSCAN++(Jang & Jiang, 2019) and VDCT(Ghaemi & Farnaghi, 2019).

5. CRITICAL REVIEW OF IMPROVED MODELS OF DBSCAN

5.1 VARIED DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

This model addresses the challenge of clustering varied density dataset. For a dataset generated similar to the one depicted in Fig.1a, the equivalent k-distplot is depicted in Fig.1b. As we observe the k-distplot, it is easy to decide the ε value as 3. But the figures depicted in Fig.2a. and Fig.2b. tells us a different story. K-distplot depicted in Fig.2b. has three steep variations in the densities. So, one ε value will give an inefficient clustering model as shown in Fig.2a.

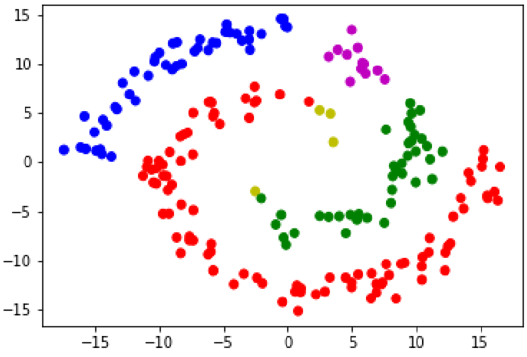

Fig.1a. Set of data points

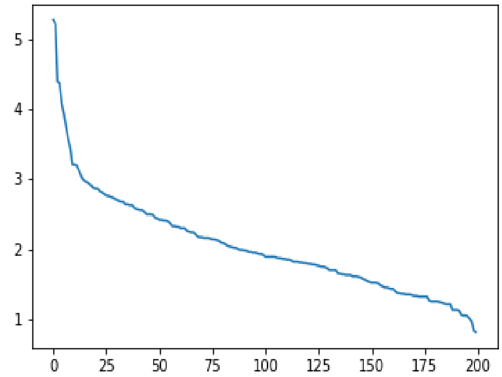

Fig.1b. K-distplot

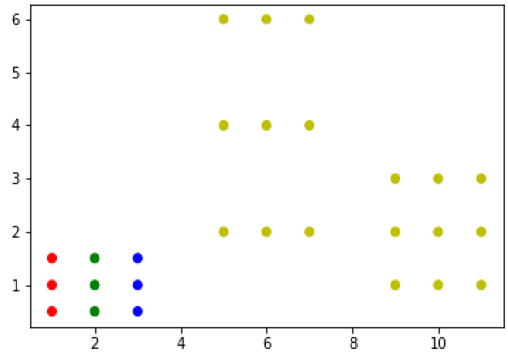

Fig.2a. Set of data points

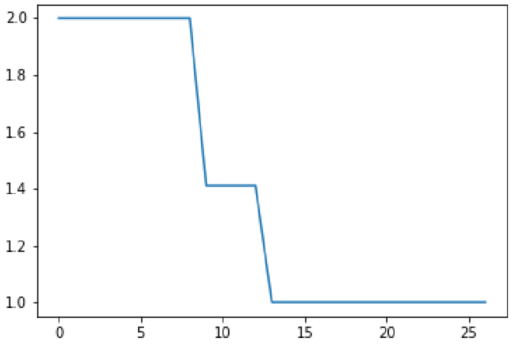

Fig.2b. K-distplot

From the figures above, it is clearly evident that, density variation is not taken into account in DBSCAN. To overcome this problem, VDBSCAN (Peng et al., 2007) was proposed. The basic idea of the algorithm is that a set of ε values are extracted through k-distplot according to various densities. With multiple ε values, clusters of varied density are identified. For every stage of clustering, the points that are already marked for forming clusters are ignored. The steps for finding ε of VDBSCAN are as follows.

Selection of ε values using K-distplot. K-distplot is used to analyse the density levels of the dataset. It efficiently identifies the steep variation observed in the density levels of the dataset. Adapt DBSCAN for larger ε value. Some clusters are formed based on this parameter and some points are identified as noise. Adapt DBSCAN for the next smaller ε value. But this time the points marked as noise are only taken for consideration. Repeat the process for all the values of chosen ε values.

Advantages: This algorithm addresses the issue of spatial heterogeneity in the data with the help of different parameters for detecting clusters in an area based on the density in that area.

Disadvantages: This algorithm deteriorates when used in high dimensional datasets. So, for a dataset derived from social media like twitter, this algorithm slows down due to the high dimensional nature of dataset.

5.2 AUTOMATIC GENERATION OF EPS FOR DBSCAN

AGED takes into account two challenges of DBSCAN –i) calculating ε value and ii) handling varied density datasets. AGED points out that estimation of *MinPts* can be done easily as the value is discrete but the estimation of ε value is continuous so it needs some technique to generate the value. To demonstrate the first challenge i.e. how a small change in ε value affects the clustering, the authors have taken spiral dataset, shown the k-distgraph in Fig.3a and the clustering results for ε = 1.23 and ε =1.0 with *MinPts* 3 in the Fig.3b. and Fig.3c. respectively.

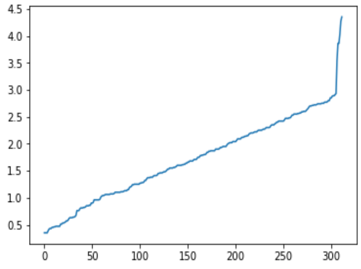

Fig.3a. K-distplot

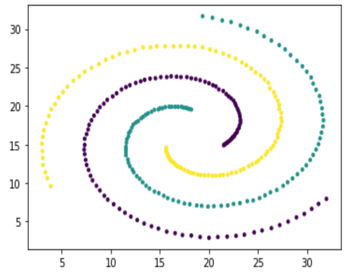

Fig.3b. DBSCAN with ε = 1.23

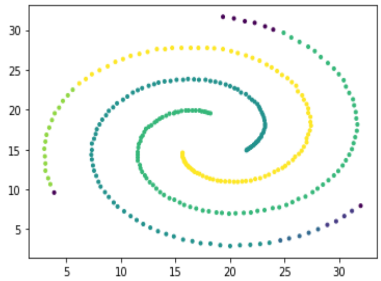

Fig.3c. DBSCAN with ε = 1.0

To demonstrate the second challenge – i.e., handling varied density dataset, we have shown a varied density dataset and how DBSCAN cannot identify different density is also shown in the Fig.4a, Fig.4b. and Fig.4c. From the figures shown below, it is very evident that DBSCAN is unable to catch the pattern of different density dataset. Fig.4b. shows the clustering with ε Value of 0.05 and Fig.4c. shows the clustering with ε value of 0.09. In Fig.4b., only one cluster is identified. In Fig.4c., two different density clusters are identified as one cluster.

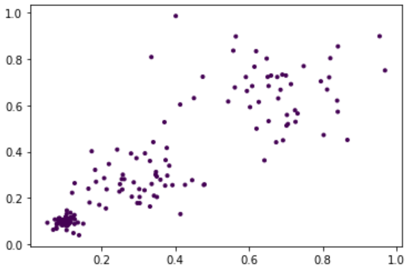

Fig.4a. Varied Density Dataset

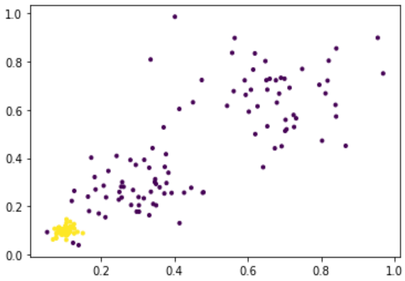

Fig.4b. DBSCAN with a smaller ε value

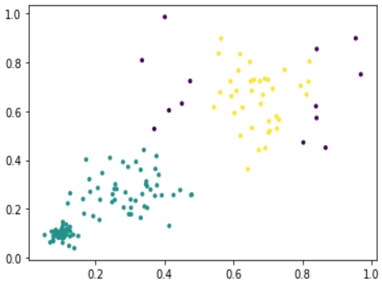

Fig.4c. DBSCAN with a bigger ε value

To tackle these two challenges, AGED model generates multiple ε values out of which one value is chosen for uniform density dataset. For varied density dataset, this process is done in iterations. While the whole dataset is considered in the first iteration, only the noise points identified in the first iteration is considered for the second iteration and so on. Calculation of ε is done through min-max normalization and binning concepts.

Advantages: This paper is one of the promising studies for automatically calculating the input parameters in density-based clustering which directly impacts the clustering quality

Disadvantages: Though this algorithm generates ε values automatically, it still takes the other parameter – *MinPts* from the user. Selection of *MinPts* on varied density dataset is a challenge in itself. This algorithm considers only 2-dimensional data for implementation.

5.3 ADAPTIVE DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

Adaptive DBSCAN analyses the k-dist graph through visual inspection and marks the steep change in the slope. The points of steep change are the ε values. For each ε value, the corresponding *MinPts* is also calculated mathematically by adding the number of points present ε-neighborhood for each data points and dividing it by number of data points.

Advantages: This paper takes in only the dataset as input. The parameters of density-based clustering are calculated mathematically by considering different parameters for different density regions present in the data.

Disadvantages: This study does not compare the performance of the proposed model with the based density-based clustering model. So, the effectiveness of the proposed model is not known. It implements the model only on one dataset. So, the actual performance cannot be decided for this model. There is no metric used to evaluate the proposed model. So, evaluation cannot be quantified.

5.4 AE-DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

AE-DBSCAN algorithm takes in the dataset and one parameter of density-based clustering *MinPts* as inputs. The paper proposes three strategies to calculate the ε value. It uses k-distgraph to find the slope after sorting the k-distvalues for all the strategies. In the proposed strategy, the slope above μ(slope) + σ(slope) is considered and the relevant value in k-distis taken as ε. This paper compares the performance of the proposed method with two other strategies of finding the ε value - Strategy1: $\varepsilon = \mu(\text{slope}) + 2\sigma(\text{slope})$ and Strategy2: $\mu(\text{slope}) - \sigma(\text{slope}) < \varepsilon < \mu(\text{slope}) + \sigma(\text{slope})$.

Advantages: This paper proposes a statistical way of finding the value of one of the parameters of density-based clustering.

Disadvantages: Though this algorithm generates ε value automatically, it still takes the other parameter – *MinPts* from the user. This algorithm considers only 2-dimensional data for implementation. Also, the performance of this algorithm is not compared with any other density clustering model. So, the actual impact of this model with respect to other models is unknown.

5.5 DENSITY BASED SPATIAL TEXTUAL CLUSTERING

DBSTexC algorithm handles content heterogeneity by taking into account one more parameter *MaxPts*, along with ε and *MinPts*. The interpretation of ε and *MinPts* are same as mentioned in DBSCAN. *MaxPts* refers to maximum number of points of interest irrelevant tweets. So, the core point should not only have *MinPts* within ε neighborhood but also should be restricted to *MaxPts* number of irrelevant tweets. Then only, it is eligible for forming the cluster. To demonstrate the above-mentioned point, consider the set of data points shown in the Fig.5. As we observe the figure, there are three types of data which are indicated in blue (representing a particular Point of Interest), in red (representing another Point of Interest). The one which is represented in black stars represent the irrelevant points of interest. While identifying the patterns, DBSCAN misses this aspect and clusters according to the spatial proximity. The DBSTexC algorithm takes into consideration these points while discovering the clusters.

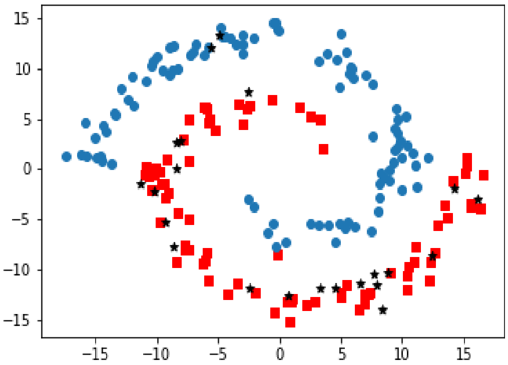

Fig.5. Set of data points with Point of interest relevance shown in red and blue and points of irrelevance shown in black

For implementing the algorithm, the authors have used twitter API to extract tweets containing tweet content and its geo-spatial information – latitude and longitude. Query searching based on point of interest has been done on the data to extract tweets. Finally, dataset will have subsets of points of relevance and points of irrelevance. Number of relevant points should be greater than or equal to *MinPts* and number of irrelevant points should be lesser than or equal to *MaxPts*. ε denotes the radius of the query region.

This algorithm takes into account one more parameter MaxPts, which results in discovering refined clusters. This algorithm handles the textual input also for clustering in addition to geographical coordinates.

Advantages: This paper discovers a way to extend DBSCAN by considering both POI-relevant tweets and POI-irrelevant tweets and then evaluate the clustering performance of DBSTexC algorithm while showing its superiority over DBSCAN. They have also analysed the computational capacity.

Disadvantages: As there is one extra parameter in this study, it results in refined clusters. But it increases the time complexity of the algorithm to O(n^2+nm), where n is the size of the dataset and m is the subset of the dataset.

5.6 DBSCAN++: TOWARDS FAST AND SCALABLE DENSITY CLUSTERING

DBSCAN++ is the proposed algorithm in this paper which is a step towards an improved version of DBSCAN. DBSCAN++ is based on the notion that there need to be only a subset of data points for which the density estimates need to be computed. The authors have proposed two strategies to choose these points - uniform and greedy K-center-based sampling. In the first strategy, it selects a subset of data points from a uniformly sampled points from the entire dataset. This subset of points is used to compute the density. From that subset, it identifies the core point and builds the neighbourhood graph. In the second strategy, it uses K-center to find the subset points.

This paper runs in a small amount of time compared to DBSCAN, while giving optimum performance and uniformly producing good clustering options across varied hyper-parameter settings. For the detection of outliers, it delivers similar results of DBSCAN

Advantages: This is an important study which makes an attempt to reduce the run-time complexity of density-based clustering.

Disadvantages: This study does not address the issue of varied-density datasets. Both the parameters of density-based clustering are taken as input from the users which will affect the quality of clustering.

5.7 A VARIED DENSITY-BASED CLUSTERING APPROACH FOR EVENT DETECTION FROM HETEROGENEOUS TWITTER DATA

The authors of (Ghaemi & Farnaghi, 2019) have proposed an algorithm called VDCT (Varied Density-based spatial Clustering for Twitter data), as an extended version of VDBSCAN, to locate and collect geo-tagged events from Twitter data where there is varied density of data. In addition to cluster data of arbitrary shape with no prerequisite knowledge on the cluster count and by efficiently handling noise, VDCT is able to detect clusters of varied density. As our study is focused on spatial clustering, we discuss the pros and cons of this algorithm with respect to only spatial clustering.

Advantages: The proposed algorithm has addressed the following challenge which other clustering algorithms encounters – other versions use k-dist graph to address the varied density data points. But this method is efficient only for small amount of data points. This algorithm works efficiently on large datasets of Twitter data. This uses exponential spline interpolation, in place of k-distgraph for addressing the varied density data points.

Disadvantages: Both the parameters of density-based clustering are taken as input from the users which will affect the quality of clustering.

6. COMPARISON OF VARIOUS ASPECTS OF ALGORITHMS

6.1 DATASETS REVIEW

DBSCAN algorithm uses Synthetic sample databases and the database of the SEQUOIA 2000 benchmark data having 4 attributes and 62,584 instances and its worst-case run-time complexity is O(n^2). VDBSCAN uses Synthetic database with 2-dimensional data and its worst-case run-time complexity is O(n^2). AGED algorithm uses standard multi-dimensional Datasets like Spiral, Wine, Iris, Compound, Zoo, Canme for evaluating the model and its run-time complexity is not mentioned in the paper. DBSTexC makes use of Geotagged Tweets extracted from Twitter API containing (Text, Lat, Lon) and its worst-case run-time complexity is O(x^2+xy) where x and y denote the Point of interest relevant and irrelevant tweets respectively.

6.2 PERFORMANCE METRICS

DBSCAN uses Visual Inspection and the model is compared with CLARANS(Raymond T. Ng and Jiawei Han, 2002) model. VDBSCAN and DBSCAN++ compare its performance with DBSCAN. AGED model uses Accuracy, Silhouette score, Dunn Index, Entropy and Pearson Gamma as performance metrices and its performance is compared with DBSCAN. ADAPTIVE-DBSCAN and AE-DBSCAN do not compare its implementation with any other models. DBSTexC uses F1 score as the performance metric and its comparison is done with DBSCAN. VDCT is compared with VDBSCAN algorithm.

The following Table.3. shows the datasets and metrics that were used in the algorithms considered for this study.

Table.3. Datasets and Metrics used in the algorithms

Algorithm	Datasets Used	Performance Metric	Runtime Complexity
DBSCAN	Standard clustering Datasets and the database of the SEQUOIA 2000 benchmark data	Visual Inspection	$O(n^2)$
VDBSCAN	Synthetic dataset with 2-dimensional data	No metric used	$O(n^2)$
AGED	Standard clustering Datasets like Spiral, Wine, Iris, Compound, Zoo, Canme	Accuracy, Silhouette score, Dunn Index, Entropy and Pearson Gamma	Not mentioned
ADAPTIVE-DBSCAN	Student Performance Dataset for clustering based on performance	No metric used	Not mentioned
AE-DBSCAN	2D Datasets - Compound, Complex9, R15	Accuracy	Not mentioned
DBSTexC	Geotagged Tweets	F1 score	$O(n^2+nm)$
DBSCAN++	Standard clustering Datasets like Wine, Iris, Spam, Zoo, MNIST	Adjusted RAND index and Adjusted Mutual Info Score	O(nm)
VDCT	Geotagged Tweets	Davies-Bouldin, Silhouette score and Dunn Index	Not mentioned

7. CONCLUSION AND FUTURE WORK

This paper illustrates the working of DBSCAN and compares with other three algorithms. While VDBSCAN addresses the varying density issue, AGED, AE-DBSCAN and ADAPTIVE DBSCAN algorithm calculate density parameters automatically. DBSTexC and VDCT focus on identifying quality clusters by considering textual information also. DBSCAN++ aims at reducing the run-time complexity. Based on the study, we can observe that there have been enough studies done on calculating ε values. But there has not been any study that has handled the other important parameter which is the *MinPts*. In our future work, we focus on a way to calculate both ε and *MinPts* based on the density of data and adapt the technique of DBSCAN++ to reduce the run-time complexity. Our proposed model will be evaluated thoroughly by comparing it with the original paper using the standard clustering metrics to ensure the efficacy of the proposed method. Incorporating the developed method on social media data is also considered for future study.

REFERENCES

[1] Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Record*, *27*(2), 94–105. https://doi.org/10.1145/276305.276314

[2] Ahmed, M. A., Baharin, H., & Nohuddin, P. N. E. (2020). Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses. *International Journal of Advanced Computer Science and Applications*, *11*(8), 248–254. https://doi.org/10.14569/IJACSA.2020.0110832

[3] Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient Machine Learning for Big Data: A Review. *Big Data Research*, *2*(3), 87–93. https://doi.org/10.1016/j.bdr.2015.04.001

[4] Ali, T., Asghar, S., & Sajid, N. A. (2010). Critical analysis of DBSCAN variations. *2010 International Conference on Information and Emerging Technologies, ICIET 2010*, *October 2015*. https://doi.org/10.1109/ICIET.2010.5625720

[5] Alraba, Y., & Al-refai, M. (2018). *Data clustering algorithms : A second look*. *4*(4), 1081–1083.

[6] Anitajesi, & Arumaiselvam; (2018). Study of Clustering Methods in Data Mining. *International Journal of Data Mining Techniques and Applications*, *7*(1), 55–59.

[7] Ankerst, M., Breunig, M. M., Kriegel, H., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *ACM SIGMOD Record*, *28*(2), 49–60.

[8] Bhattacharjee, P., & Mitra, P. (2021). A survey of density based clustering algorithms. *Frontiers of Computer Science*, *15*(1). https://doi.org/10.1007/s11704-019-9059-3

[9] Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

[10] Bose, A., Munir, A., & Shabani, N. (2017). A Comparative Quantitative Analysis of Contemporary Big Data Clustering Algorithms for Market Segmentation in Hospitality Industry. *ArXiv*, 1–6.

[11] Ghaemi, Z., & Farnaghi, M. (2019). A Varied Density-based Clustering Approach for Event Detection from Heterogeneous Twitter Data. *ISPRS International Journal of Geo-Information*, *8*(2). https://doi.org/10.3390/ijgi8020082

[12] Harshini, G. N., & Gobi, N. (2020). Student Feedback Sentiment Analysis system for distance education using Arm with K-means clustering. *Ictact Journal on Soft Computing*, *6956*(April), 2071–2075. https://doi.org/10.21917/ijsc.2020.0294

[13] Hinneburg, A., & Keim, D. A. (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise-based Clus- tering, Clustering of High-dimensional Data, Clustering in Multimedia Databases, Clustering in the Presence of Noise. *Proceeding of the 4th International Conference on Knowledge Discovery and Data Mining(KDD)*, *c*, 58–65. https://ocs.aaai.org/Papers/KDD/1998/KDD98-009.pdf

[14] Huang, Y., Li, Y., & Shan, J. (2018). Spatial-temporal event detection from geo-tagged tweets. *ISPRS International Journal of Geo-Information*, *7*(4). https://doi.org/10.3390/ijgi7040150

[15] Jang, J., & Jiang, H. (2019). DBScan++: Towards fast and scalable density clustering. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*, 5348–5359.

[16] Jiawei, H., Micheline, K., & Jian, P. (2012). *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers.

[17] Kakkar, P., & Parashar, A. (2014). Comparison of Different Clustering Algorithms using WEKA Tool. *International Journal of Advanced Research in Technology, Engineering and Science*, *1*(2), 20–22.

[18] Kankal, S. S., Dhakne, A. R., & Tayade, Y. R. (2017). *A Brief Survey on Clustering Algorithms in Data Mining Jawaharlal Nehru Engineering College , Aurangabad , India*. *4*(11), 494–497.

[19] Lakshmi, M., Sahana, J., & Venkatesan, P. (2018). Review on Density Based Clustering Algorithms for Big Data. *International Journal of Data Mining Techniques and Applications*, *7*(1), 13–20. https://doi.org/10.20894/ijdmta.102.007.001.003

[20] Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of Massive Datasets. In *Mining of Massive Datasets*. https://doi.org/10.1017/cbo9781139924801

[21] Loh, W. K., & Park, Y. H. (2014). A survey on density-based clustering algorithms. In *Lecture Notes in Electrical Engineering: Vol. 280 LNEE* (pp. 775–780). https://doi.org/10.1007/978-3-642-41671-2_98

[22] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, X. X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Kdd96 (Ed.), *KDD-96 Proceedings. Copyright © 1996, AAAI* (p. 6). KDD96. www.aaai.org

[23] McInnes, L., & Healy, J. (2017). Accelerated Hierarchical Density Based Clustering. *IEEE International Conference on Data Mining Workshops, ICDMW*, *2017-Novem*, 33–42. https://doi.org/10.1109/ICDMW.2017.12

[24] Murugan, D., & Rathna, S. S. (2019). FUZZY BASED PRIVACY PRESERVED K-MEANS CLUSTERING. *ICTACT Journal On Soft Computing*, *6956*(October), 2011–2014. https://doi.org/10.21917/ijsc.2019.0284

[25] Naik Gaonkar, M., & Sawant, K. (2013). AutoEpsDBSCAN : DBSCAN with Eps Automatic for Large Dataset. *ISSN (Print*, *22*, 2319–2526.

[26] Nguyen, M. D., & Shin, W. Y. (2017). DBSTexC: Density-Based spatio-Textual clustering on twitter. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 23–26. https://doi.org/10.1145/3110025.3110096

[27] Ozkok, F. O., & Celik, M. (2017). A New Approach to Determine Eps Parameter of DBSCAN Algorithm. *International Journal of Intelligent Systems and Application in Engineering*, *5*(4), 247–251. https://doi.org/10.1039/b0000

[28] Pambudi, E. A., Badharudin, A. Y., & Wicaksono, A. P. (2021). Enhanced K-Means By Using Grey Wolf Optimizer for Brain Mri Segmentation. *ICTACT Journal On Soft Computing*, *11*(03), 2353–2358. https://doi.org/10.21917/ijsc.2021.0336

[29] Parimala, M., Lopez, D., & Senthilkumar, N. C. (2011). *A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases*. *31*, 59–66.

[30] Peng, L., Dong, Z., & Naijun, W. (2007). VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise. *Proceedings - ICSSSM'07: 2007 International Conference on Service Systems and Service Management*, 1–4. https://doi.org/10.1109/ICSSSM.2007.4280175

[31] Priyadarshini, Sudha, Usha, & Freeda, A. (2016). Implementation of Adaptive DBSCAN for Cluster Analysis. *IJSTE - International Journal of Science Technology & Engineering*, *2*(09), 164–168.

[32] Ram, A., Jalal, S., Jalal, A. S., & Kumar, M. (2010). A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases. *International Journal of Computer Applications*, *3*(6), 1–4. https://doi.org/10.5120/739-1038

[33] Raymond T. Ng and Jiawei Han. (2002). CLARANS - A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, *14*(5), 1003–1016. https://doi.org/10.1109/TKDE.2002.1033770

[34] Sander, J., Ester, M., Kriegel, H., Knowledge, X. X.-D. mining and, & 1998, U. (n.d.). GDBSCAN. *Springer*, *1*. https://doi.org/https://doi.org/10.1016/j.datak.2006.01.013

[35] Shi, J., Mamoulis, N., Wu, D., & Cheung, D. W. (2014). Density-based place clustering in geo-social networks. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 99–110. https://doi.org/10.1145/2588555.2610497

[36] Shi, Z., & Pun-Cheng, L. S. C. (2019). Spatiotemporal data clustering: A survey of methods. In *ISPRS International Journal of Geo-Information* (Vol. 8, Issue 3). https://doi.org/10.3390/ijgi8030112

[37] Soni, N., & Ganatra, A. (2016a). MOiD (Multiple Objects incremental DBSCAN) – A paradigm shift in incremental DBSCAN. *International Journal of Computer Science and Information Security (IJCSIS)*, *14*(4), 316–346. https://doi.org/10.6084/M9.FIGSHARE.3362410.V1

[38] Soni, N., & Ganatra, A. (2016b). AGED (Automatic Generation of Eps for DBSCAN). *International Journal of Computer Science and Information Security (IJCSIS)*, *14*(5), 536–559.

[39] Tran, T. N., Drab, K., & Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems*, *120*, 92–96. https://doi.org/10.1016/j.chemolab.2012.11.006

[40] Valarmathy, N., & Krishnaveni, S. (2019). Performance evaluation and comparison of clustering algorithms used in educational data mining. *International Journal of Recent Technology and Engineering*, *7*(6), 103–112.

[41] Viswanath, P., & Pinkesh, R. (2006). L-DBSCAN: A fast hybrid density based clustering method. *Proceedings - International Conference on Pattern Recognition*, *1*, 912–915. https://doi.org/10.1109/ICPR.2006.741

[42] Viswanath, P., & Suresh Babu, V. (2009). Rough-DBSCAN: A fast hybrid density based clustering method for large data sets. *Pattern Recognition Letters*, *30*(16), 1477–1488. https://doi.org/10.1016/j.patrec.2009.08.008

[43] Vu, D. D., To, H., Shin, W. Y., & Shahabi, C. (2016). GeoSocialBound: An efficient framework for estimating social POI

boundaries using spatio-textual information. *3rd International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, GeoRich 2016 - In Conjunction with SIGMOD 2016*, 13–18. https://doi.org/10.1145/2948649.2948652

[44] Xu, X., Ester, M., Kriegel, H. P., & Sander, J. (1998). Distribution-based clustering algorithm for mining in large spatial databases. *Proceedings - International Conference on Data Engineering*, 324–331. https://doi.org/10.1109/icde.1998.655795

[45] Zhang, T., Ramakrishnan, R., & Miron, L. (1997). BIRCH: An Efficient Data Clustering Method for Very Large Databases. *Data Mining and Knowledge Discovery*, *1*(2), 141–182. https://doi.org/10.1023/A:1009783824328