INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

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

Korra Bichya, Research Scholar and Dr. Suribabu Potnuri, Professor
Department of Computer Science and Engineering,
J.S. University, Shikohabad, U.P. India email: korra.bichya@gmail.com

ABSTRACT

Recently, there has been a boom in interest in the subject of spatial database mining, as well as research that has been concentrated on the subject. There is a significant quantity of geotagged information shown on social networking websites. It is vital to swiftly extract significant insights from this massive quantity of information in order to uncover fascinating patterns and recognize events. This is because the information is so enormous. For the purpose of managing the many types of data that are accessible on social media platforms, it is necessary to use proper approaches in order to reach the desired level of efficacy. The purpose of this research is to give a full understanding of a number of density-based clustering algorithms, as well as their specific applications in a variety of fields, the datasets that were employed, and the methodologies that were applied for data extraction. An additional key emphasis of this investigation is the assessment of the algorithms that are included in this survey.

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

1. INTRODUCTION

Over the course of the last few years, the development of new technologies and the devices that correspond to them has led to the gathering of a significant quantity of data. In light of this, it is of the utmost importance to construct sophisticated and efficient models of predictive intelligence in order to fulfill the expectations of the future. The reference to the study (Al-Jarrah et al., 2015) is given here. The recent proliferation of social media platforms has resulted in the availability of a substantial amount of data for the purpose of analysis. Several features, such as attitudes, behaviors, trends, and impacts, may be investigated with the use of these data. It is essential to make use of appropriate algorithms that are tailor-made to the particular types of data that are being evaluated in order to conduct research in an efficient manner. We may be able to discover useful insights that are generated from the information that is already available if we analyze this data. By way of illustration, by examining the profile of a user, we can discover fascinating patterns regarding the demographic information that they provide. Through the use of the textual

ISSN: 2057-5688

data, we are able to extract sentiment analysis, which is very beneficial for the study on marketing and behavior. There are a variety of formats that this information may be presented in, including images, videos, text, geocodes, and others. Through the use of unsupervised algorithms, it is possible to draw patterns from the information that is being provided. Clustering is a method that is often used in the area of unsupervised learning and is highly praised.Leskovec et al. (2014) state that

Following is an explanation of the structure that will be used for the succeeding portions of the article. In Section 2, a quick review of many clustering algorithms is presented, with a special focus placed on DBSCAN (Martin Ester, Hans-Peter Kriegel, and Jiirg Sander, 1996). The evaluation of a number of research publications that are associated with clustering algorithms is the primary topic of Section 3, with a special emphasis placed on DBSCAN specifically. When it comes to properly managing a wide variety of data sets, including high-dimensional data, it investigates the various modifications and upgrades that have been made to DBSCAN. The selection of the study work that is included in this book is discussed in Section 4, which offers a condensed explanation of the reasoning for the selection. The comparison between DBSCAN and other upgraded models is investigated in Section 5. In the sixth section, a quick review of the different features of the algorithms that were investigated in this research is presented. The most comprehensive analysis is presented in Section 7, which also includes a list of possible topics for further investigation.

2. CLUSTERING ALGORITHMS

In order to better understand clustering techniques, the following categories may be used. A number of different methods for clustering data are discussed in the article titled "Partition-Based, Hierarchical, Grid-Based, and Density-Based" that was written by Jiawei and colleagues (2012).

2.1 PARTITION-BASED METHODS

Among the many examples of partition-based clustering techniques, K-Means clustering is one of the most well-known. This particular clustering technique requires the user to enter the appropriate number of clusters as input. Following this, the dataset is partitioned into the clusters that the user has specified. In order to improve upon this clustering technique, many improvements have been implemented. An example of a model that makes use of the grey-wolf optimizer to do brain MRI segmentation can be found in Pambudi et al.'s 2021 publication. A specified number of clusters is required in this scenario, and noise does not have an impact on the performance of this strategy. It is possible to circumvent these two issues using K-Means by using density-based clustering. The study that was carried out by Murugan and Rathna (2019) investigates the consequences that are brought about by the use of a fuzzy privacy protection strategy in clustering. Harshini and Gobi (2020) investigate the use of K-Means clustering for the purpose of conducting sentiment analysis in their article.

2.2 HIERARCHICAL-BASED METHODS

One method of grouping data is called hierarchical clustering, and it uses a tree structure called a dendrogram to highlight the similarities that exist between different groups. The clusters that are included inside the dendrogram are the

ISSN: 2057-5688

same as the linked clusters that are a part of a particular dataset. There are two main types of algorithms that make up the hierarchical approach. For example, the agglomerative method. 2. The method of division. Both of these methods might be established using a dendrogram. There are a number of factors that are used by the hierarchical clustering approach in order to identify which clusters should be merged at each step. When using the agglomerative method, each point is treated as a distinct cluster, and then the clusters are progressively combined until a final cluster is established. One of the most notable improvements to the agglomerative algorithm is the BIRCH algorithm, which was created by Zhang and colleagues in 1997. When using the divided approach, the initial step is to allocate each data point to a single cluster. Subsequently, the data points are partitioned into further smaller clusters in an iterative manner until the final desired conclusion is obtained.

2.3 GRID-BASED METHODS

The data points are separated into a preset number of compartments, which results in the formation of a structure that resembles a grid. This technology's key advantage is its lightning-fast processing speed, which is purely dictated by the number of cells and is not impacted by the number of objects. This is the most significant advantage of this technology. When it comes to tackling a number of issues associated with spatial data mining, this strategy is quite effective. An instance of this method is provided by the algorithm that is often referred to as CLIQUE (Agrawal et al., 1998...).

1.1 DENSITY-BASED METHODS

The DBSCAN method and its variants, which were created by Martin Ester, Hans-Peter Kriegel, and Jiirg Sander in 1996, are very efficient when it comes to recognizing clusters that have irregular forms and dealing with noisy data. It is not necessary to have previous information of the number of clusters that are to be described in order to use these techniques. This study is concentrating primarily on the density-based clustering approach because of its adaptability and the fact that it continues to be popular in a wide variety of applications.

3 RELATED WORKS

Different review papers have been written on Clustering Algorithms in general, with a particular emphasis on Density-based clustering. These publications have been published.

The details about different review papers are tabulated below in Table.1.

Table.1. Different Review Works on Clustering Algorithms

ISSN: 2057-5688

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 an d 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

ISSN: 2057-5688

Despite the fact that these surveys include a variety of clustering methods, none of them have specifically focused on the current developments in density-based clustering. The purpose of this study is to investigate the operation of seven studies that are founded on the density-based clustering approach that has been widely used in the past.

Numerous techniques have been developed by researchers in order to extract useful information from data sets that are very large. A good illustration of such an algorithm is the density-based clustering algorithm. Martin Ester, Hans-Peter Kriegel, and Jiirg Sander's DBSCAN method, which was developed in 1996, is the first algorithm to implement density-based clustering.

3.1 REVIEW ON DBSCAN

This research not only investigates the operation and results of existing models that are associated with density-based clustering, but it also carries out an exhaustive analysis of the efficacy and efficiency of these models on a variety of datasets.
The core idea behind density-based clustering is that every data point inside a cluster must have a neighborhood with a radius ε greater than zero that includes a minimum number of data points. This is the underlying principle behind density-based clustering. More specifically, the amount of data points in the area must be larger than a certain threshold in order to meet the requirements.

3.2 EVOLUTION OF DBSCAN ALGORITHM

Sander et al. (n.d.) created GDBSCAN, which is an improvement on the traditional DBSCAN algorithm. This improvement is achieved by extending the concept of neighborhood beyond the traditional E-neighborhood technique. The "cardinality" of the neighborhood is also determined using various approaches, which are presented in this article. (Viswanath & Pinkesh, 2006) proposes a novel strategy that makes use of two distinct sorts of prototypes. One of these prototypes is designed to reduce the amount of time that is required, while the other is centered on reducing the amount of deviation that occurs in the final product. Through the use of the leader clustering methodology, prototypes are produced. A two-tiered hierarchy with unique threshold values is produced as a consequence of this decision. Through this study, a hybrid clustering approach that is scalable was presented for the purpose of creating density-based clusters of any certain shape. During the first phase, you will be responsible for developing two prototypes by using the leaders clustering concept. The prototypes are used by the traditional DBSCAN. It has been shown that the l-DBSCAN strategy, which was recommended, functions more effectively than the DBSCAN method when applied to the whole dataset. Rapid clustering was used by the authors of (Viswanath & Pinkesh, 2006) in order to get prototypes that were referred to as leaders in his study (Viswanath & Suresh Babu, 2009). The following enhancements have been made by the writers in order to improve their own rendering of previous studies: Delineating the features of leaders from a theoretical standpoint and providing an explanation of how the proposed strategy carries out its functions, In order to conduct an analysis of the proposed model, rough set theory was considered. In contrast to the running time of DBSCAN, which has been shown to be quadratic, the running time of crude

ISSN: 2057-5688

DBSCAN has been found to be linear. The difficulty of applying density-based clustering to data with various densities is addressed in the research conducted by Peng et al. (2007), which investigates datasets that include a variety of densities. The approach that is described in (Tran et al., 2013) serves to reassess the DBSCAN concept and to make adjustments in order to improve its overall performance. The purpose of this research is to propose a method that is capable of managing both geographical and non-spatial data kinds. Additionally, it addresses the issue of border data points that are allocated to nearby clusters using this method. The technique makes use of core-density-reachable chains, which are used instead of the more traditional density-reachable objects since they only take into consideration core points. There is a correlation between the border object's membership in the core-density-reachable chain and the cluster to which it is allocated. The core principles of the DBSCAN algorithm are maintained by this method, which also provides the opportunity to improve clustering results by addressing the issue of border items at the same time. The research that is referred to in (Naik Gaonkar & Sawant, 2013) makes use of an automated technique to develop input parameters that are capable of properly identifying clusters that have various densities respectively. An incremental DBSCAN approach that is capable of processing multiple data items concurrently is referred to as MOiD (Multiple items incremental DBSCAN), and it is introduced in the methodology that is detailed in Soni and Ganatra's 2016a publication. Soni and Ganatra (2016b) presented a mathematical method for obtaining the ε value, which is one of the input elements inside the Density-based clustering technique. This method was introduced in their academic paper. The K-nearest neighbor analysis protocol is used in this technique. Priyadarshini et al. (2016) established a technique that could calculate the parameters of density-based clustering. This approach was created by the authors of the study. The authors (Ozkok & Celik, 2017) have proposed a model that incorporates a technique known as AE-DBSCAN. This method is used to automatically calculate the value of the neighborhood radius. The model makes use of the k-distgraph in order to discover the gradients that are present within the density of the dataset. A calculation is made to determine the average and standard deviation of all slopes that are not zero. After that, it detects the first gradient that surpasses both the mean and the standard deviation, and it assigns that particular gradient the designation of the ε value. A research that was carried out in 2017 by Nguyen and Shin sheds light on the shortcomings of DBSCAN and the many modifications that it has undergone. It has been discovered that the region around a Point of Interest is often made up of points that either have or do not have annotated Point of Interest phrases. These points are referred to as Point-relevant points and Point-irrelevant points, respectively. During the clustering process, however, DBSCAN only takes into consideration the points that are important to the cluster. In order to find a solution to this problem, a new method known as DBSTexC has been devised. In order to carry out density-based clustering on Twitter, this technique integrates text data with the DBSCAN model. In order to do this, it is designed to remove geographically concentrated locations that include a significant amount of geotagged messages that are not relevant. The HDBSCAN algorithm, which was developed by McInnes and Healy (2017), was first implemented by first establishing a hierarchy in density-based clustering and then picking clusters according to stability criteria. Because of

ISSN: 2057-5688

the availability of noisy data, it is possible that erroneous clusters may be produced if the analysis is limited to a particular kind of input data, such as tweets that are associated with a place of interest (POI). In the research carried out by Jang and Jiang (2019), the model known as DBSCAN++ was provided as a solution to the issue of the amount of processing time that DBSCAN requires. The research carried out by Huang et al. (2018) and Ghaemi & Farnaghi (2019) includes the use of density-based clustering for the purpose of event detection. According to the research conducted by Shi et al. (2014) and Vu et al. (2016), the use of density-based clustering is utilized on social network data for a variety of applications.

4. MOTIVATION

According to the findings of the literature review, it is clear that there is a dearth of research or surveys concerning the improved iterations of density-based clustering algorithms. There are seven algorithms that we have discovered that are modified versions of DBSCAN that explicitly address the deficiencies that it has. Detailed explanations of the rationale for the selection of the algorithms are included below. Varied Density Based Spatial Clustering of Applications with Noise is what an acronym known as VDBSCAN stands for.The model that was developed by Peng et al. (2007) outlines the limits of the density-based approaches that are currently in use in terms of properly recognizing all important clusters in datasets that have different densities. When many values of ε are taken into consideration, it becomes an easy process to accurately detect clusters that possess different densities concurrently. The approach known as AGED (Automatic Generation of ε for DBSCAN), which was introduced by Soni and Ganatra in 2016, is designed to overcome two significant challenges that are faced by traditional density-based algorithms. These challenges include successfully managing datasets that include changing densities and precisely calculating the input parameters. The findings of this study underline the fact that the density-based technique is very sensitive to even the most minute changes in the values of the input parameters they are applied to. In this work, a mathematical approach is shown for obtaining the ε value, which is one of the input parameters for density-based clustering. This method utilizes K-nearest neighbor to accomplish this determination. For the purpose of determining separate ε values for different densities, this article employs min-max normalization, which is then followed by binning process. The Adaptive DBSCAN algorithm, which was presented in the research conducted by Priyadarshini et al. (2016), offers a technique for finding the ε value and the MinPts value that is associated with it. The methodology known as AE_DBSCAN, which was first presented by Ozkok and Celik in 2017, encompasses three distinct methods for calculating the value of ε. When conducting an analysis of the region surrounding a Point of Interest, the research conducted by Nguyen and Shin (2017), which was titled Density-Based Spatial-Textual Clustering on Twitter (DBSTexC), discovered that there are geo-tags that contain keywords related to the Point of Interest as well as geo-tags that do not contain any keywords related to the Point of Interest. These geo-tags are referred to as POI-relevant and POI-irrelevant geo-tags, respectively. The clustering method of DBSCAN, on the other hand, only takes into consideration geo-tags that are relevant to POIs. This particular approach takes into account both relevant areas of interest and sets a threshold limit on any superfluous information that could be present within

ISSN: 2057-5688

the region of the ε value. This results in an improvement in clustering. Towards efficient and scalable density clustering (Jang & Jiang, 2019) is based on the concept that it is only required to produce density estimates for a subset of data points. This is the foundation upon which the DBSCAN++ algorithm is constructed. Two methods, namely greedy K-center-based sampling and uniform sampling, have been proposed by the authors as potential techniques to picking these locations. This strategy reduces the number of data points that need to be assessed during the calculation, which in turn reduces the amount of time required for the execution. For the goal of clustering, the authors of the VDCT model, which was reported in the research carried out by Ghaemi and Farnaghi (2019), used a wide variety of Twitter data that jhad a variety of features. When conducting their clustering study, the authors made use of both geographical and textual information. The functionality of the original DBSCAN algorithm is analyzed in this paper, and it is compared to seven improved versions of the algorithm. These versions are as follows: 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

The challenge of grouping datasets with different densities is an issue that may be addressed by this technique. A k-distplot that is equivalent to the one displayed in Figure 1a is used to illustrate the dataset that was collected in a way that is comparable to the one shown in Figure 1a. Upon careful examination of the k-distplot, it becomes obvious that the suitable value for the coefficient ε is 3.
The statistics shown in Figure 2a and Figure 2b, on the other hand, provide a narrative that is opposing to one another. With regard to the density, the K-distplot that is shown in Figure 2b demonstrates three separate peaks. Because of this, a single ε value will lead to an inefficient clustering model, as seen in
Figure 2a.

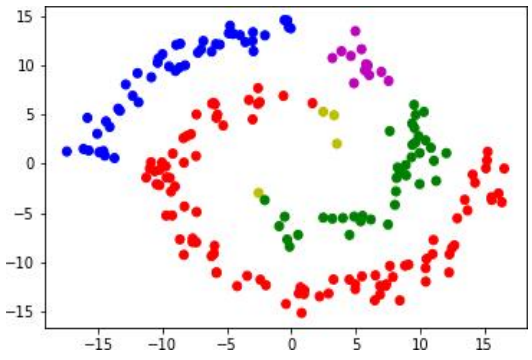

Fig.1a. Set of data points

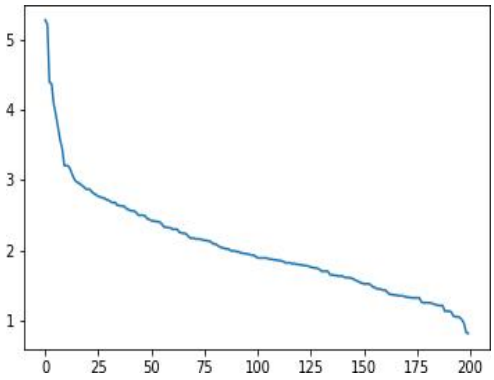

Fig.1b. K-distplot

ISSN: 2057-5688

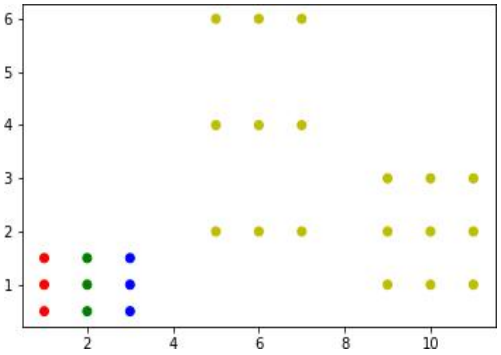

Fig.2a. Set of data points

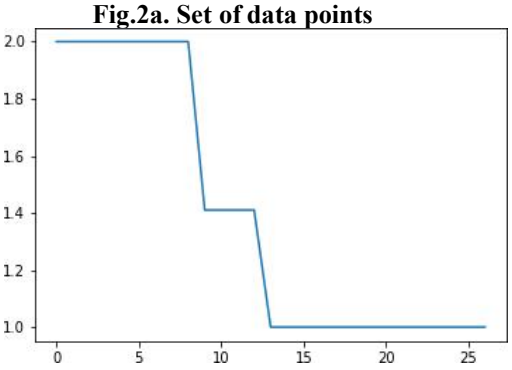

Fig.2b. K-distplot

It is abundantly obvious from the figures shown above that DBSCAN does not take density fluctuations into consideration. In order to solve this problem, a technique known as VDBSCAN (Peng et al., 2007) was developed and implemented. In order to function, the technique involves the extraction of a collection of ε values from a k-distplot, while taking into account various densities. By using a multitude of ε values, it is possible for us to recognize clusters that possess varying densities. During each stage of the clustering process, the points that have already been identified for cluster formation are ignored. Here is an overview of the technique that may be followed in order to determine the ε value of VDBSCAN.

When selecting ε values, the K-distplot is used for the selection process. When doing an analysis of the density distribution of the dataset, the K-distplot function is brought into play. The abrupt changes in the density levels of the dataset are correctly detected by it because of its accuracy. The DBSCAN algorithm should be modified so that it can tolerate a larger value for ε. When this feature is taken into consideration, clusters are produced, while some points are categorized as noise. Make adjustments to the DBSCAN algorithm so that it can handle a lower value of ε. But this time, just the points that have been recognized as being noise are being taken into consideration. Repeat the method for each of the ε values that have been chosen.
Advantages: This method addresses the issue of geographic heterogeneity in the data by using a number of features to identify clusters in a certain region, with the density of those clusters being determined by the density of specific areas within that region.

Negative aspects include the fact that the performance of this method suffers when it is used to datasets that have a significant number of dimensions. This algorithm's performance decreases when it is applied to a dataset gathered from social media sites like Twitter. This is due to the fact that the dataset contains a huge number of dimensions.

5.2 AUTOMATIC GENERATION OF EPS FOR DBSCAN

One of the most important challenges in DBSCAN is establishing the ideal ε value, and the other is properly handling datasets with different densities. AGED tackles both of these challenges. It is brought to the attention of AGED that the estimation of MinPts may be simple due to the fact that its value is discrete. Nevertheless, the estimate of the value of ε is continuous, which necessitates the use of a procedure in order to generate the value. The authors of the paper have used a spiral dataset in order to provide an illustration of the first challenge, which entails analyzing the influence of a

ISSN: 2057-5688

little alteration in the ε value on the clustering process. In Figure 3a, they have shown the k-distgraph, and in Figures 3b and 3c, respectively, they have presented the clustering results for ε values of 1.23 and 1.0, both of which have a MinPts of 3.For the purpose of illustrating the second challenge, which is dealing with datasets that have different densities, we have presented a dataset that has varying densities. Furthermore, as can be seen in Figures 4a, 4b, and 4c, we have proved that DBSCAN is unable to differentiate between different densities. The statistics that have been shown make it very evident that DBSCAN is unable to accurately capture the pattern of a dataset that has changing densities. For the sake of illustration, Figure 4b represents the clustering process with an ε value of 0.05, whereas Figure 4c demonstrates the clustering process with an ε value of 0.09. In Figure 4b, we can see that there is just one cluster that has been identified. Figure 4c has an error in classification that incorrectly identifies two separate density clusters as a single cluster.

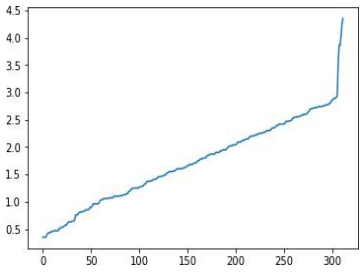

Fig.3a. K-distplot

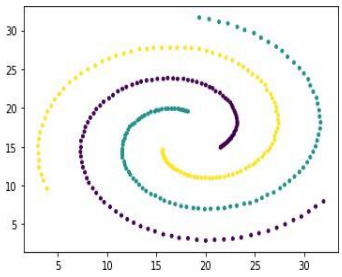

Fig.3b. DBSCAN with ε =1.23

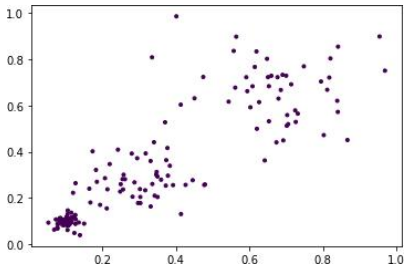

Fig.4a. Varied Density Dataset

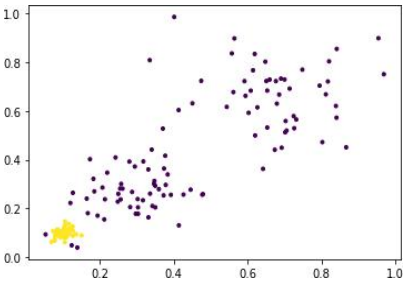

Fig.4b. DBSCAN with a smaller ε value

With the intention of addressing these two issues, the AGED model creates a number of ε values, from which a single value is selected for the uniform density dataset. In the case of datasets with varying densities, this technique is carried out in iterations. During the first iteration, the whole dataset is taken into consideration; however, during the second iteration, only the noise spots that were discovered during the first iteration are taken into consideration, and so on. When calculating ε, the min-max normalization and binning ideas are used to accomplish the calculation.

On the other hand, this work is one of the promising research for automatically computing the input parameters in density-based clustering, which has a direct influence on the quality of the clustering.

Disadvantages: Despite the fact that this technique automatically creates ε values, it still requires the user to provide the MinPts parameter. The process of selecting MinPts from a dataset with varying densities is a difficulty in and of

ISSN: 2057-5688

itself. In terms of implementation, this method only takes into consideration data in two dimensions.

5.3 ADAPTIVE DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

An adaptive DBSCAN is used to do a visual examination of the k-dist graph, which then identifies the sudden change in gradient that is present in the graph. The values of ε provide information about the particular locations in which the phenomenon takes place during a rapid and considerable fluctuation. In order to ascertain the MinPts value for every ε, the theoretical approach involves accumulating the total number of points that are located within the ε-neighborhood of every data point, and then dividing this total by the entire number of data points.

Advantages: For the purposes of this work, the dataset acts as the only input. Through the use of mathematical calculations, it is possible to ascertain the parameters of density-based clustering using this method. The calculations in question investigate a number of features that are unique to certain density zones included within the data.

6 In this particular research, there is a lack of a comparison analysis between the performance of the suggested model and the performance of the density-based clustering model that it is created upon. This is a limitation that the study must accept. Whether or if the model that was proposed is effective is yet unknown. A single dataset is the only one that the model is applied to. As a result, it is not feasible to arrive at a conclusive assessment of the performance of this model. The proposed model is evaluated without the use of any criterion. As a consequence of this, evaluation cannot be quantified using numerical numerical values. .

6.2 AE-DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

In order to perform density-based clustering, the AE-DBSCAN technique needs the dataset as well as the MinPts parameter as sources of information. In the research, three different approaches are shown for calculating the ε value. When the k-distvalues for all of the different methods are sorted using the k-distgraph, the slope can be calculated. Based on the approach that has been provided, our attention is directed on the slope that is higher than the mean slope (μ(slope)) in addition to the standard deviation of the slope (σ(slope)). The particular value that corresponds to the k-distribution is then determined to be ε. Through the comparison of the recommended strategy to two alternative approaches for obtaining the ε value, this study endeavors to assess the usefulness of the aforementioned methodology. The first method, which is known as Strategy1, involves calculating the value of ε by adding the mean slope (μ) and twice the standard deviation

ISSN: 2057-5688

of the slope (σ). The second strategy, which is known as Strategy2, takes into account the range of the mean slope minus the standard deviation (μ - σ) and the mean slope plus the standard deviation (μ + σ). This method establishes the value of ε between these two levels. Advantages include: The purpose of this study is to propose a statistical method for calculating the value of a parameter in density-based clustering.

One of the drawbacks of this approach is that while it generates the ε value automatically, it still needs the user to provide the MinPts parameter. When it comes to its execution, this algorithm only makes use of data that is two-dimensional. Moreover, the performance of this technique has not been evaluated in comparison to the performance of any other density clustering model or algorithm. There is still a lack of clarity on the specific effect that this model has in compared to other models.

6.3 DENSITY BASED SPATIAL TEXTUAL CLUSTERING

A solution to the problem of heterogeneity in content is provided by the DBSTexC approach, which incorporates an extra parameter called MaxPts in addition to the parameters ε and MinPts. Within the context of this particular scenario, the meaning of the variables ε and MinPts is in accordance with the definitions that are supplied in the DBSCAN algorithm. The maximum number of tweets that are entirely unrelated to places of interest is denoted by the symbol MaxPts. In order to fulfill the basic criterion, it is necessary for the core point to possess both MinPts inside the ε neighborhood and to be restricted to a maximum of MaxPts tweets that are irrelevant. When this occurs, it becomes ideal for the creation of clusters. As an illustration of the problem that was addressed before, let's have a look at the data points that are shown in Figure 5. Upon closer inspection of the visual, we are able to recognize three separate types of data. In order to distinguish between these categories, the color blue, which stands for a certain Point of Interest, and the color red, which stands for another Point of Interest, are used. The sites of interest that are not very significant are represented by the black stars. When it comes to detecting patterns, DBSCAN does not take this factor into consideration and instead clusters based on the proximity of geographic locations. When it comes to determining clusters, the DBSTexC algorithm takes these considerations into account.

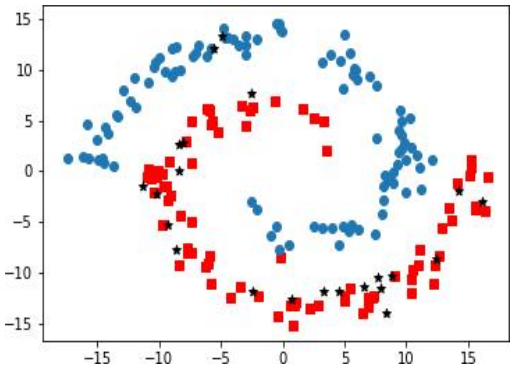

Fig.5. Set of data points with Point of interest relevance shown in red and blue and points of irrelevance shown in black

By using the Twitter Application Programming Interface (API), the authors were able to extract tweets that included not only the message's content but also its geographical information, namely its

ISSN: 2057-5688

latitude and longitude. Through the use of a query search constructed on the basis of locations of interest, tweets have been retrieved from the data. After all is said and done, the dataset will be made up of subsets that include points that are important as well as those that are irrelevant. The number of points that are relevant must be more than or equal to the minimum number of points, while the number of points that are not relevant must be less than or equal to the maximum number of points. ε is a symbol that denotes the radius of the region that is being queried. MaxPts is an extra parameter that is included into this technique, which ultimately results in the identification of clusters that are more precise. This approach not only is capable of processing geographical coordinates, but it also manages textual input for the purpose of clustering.

Advantages: An addition to the DBSCAN approach is proposed in this paper, which includes tweets that are related to POI as well as tweets that are irrelevant to POI. After that, the clustering performance of the DBSTexC method is examined, and the results demonstrate that it is superior than the DBSCAN algorithm that was first developed. In addition to that, they have investigated the possible computer capabilities.

Disadvantages: As a result of the incorporation of an extra parameter into this investigation, the clusters that were produced showed an increase in precision and quality. Nevertheless, the method's time complexity is O(n^2 + nm), where n is the size of the dataset and m is the subset of the dataset. This is the outcome of the method's implementation.

.

6.3 DBSCAN++: TOWARDS FAST AND SCALABLE DENSITY CLUSTERING

The article presents DBSCAN++, a technique that is designed to improve upon the DBSCAN algorithm that was first developed. On the basis of the premise that it is only required to construct density estimates for a subset of data points, DBSCAN++ is able to perform its operations. Two methods, namely greedy K-center-based sampling and uniform sampling, have been proposed by the authors as potential techniques to picking these locations. One method includes picking a subset of data points from the whole dataset in such a way that they are distributed uniformly over the entire set. It is via the use of this specific collection of data points that the density is determined. This subset is used to find the central point, and the neighborhood graph is constructed using that information. Identifying a selection of places is accomplished by the use of the K-center algorithm in the second method.

When compared to DBSCAN, this research demonstrates a much lower runtime, while simultaneously achieving optimum performance and reliably producing high-quality clustering alternatives over a wide range of hyper-parameter combinations. When it comes to finding outliers, it offers results that are equivalent to those of DBSCAN. A reduction in the computational complexity of density-based clustering is one of the benefits that this study intends to bring about.

One of the limitations of this study is that it does not address the issue of datasets that have changing densities. Because the users contribute input for both parameters of density-based clustering, the quality of the clustering is directly impacted by the information they supply.

6.4 A VARIED DENSITY-BASED

ISSN: 2057-5688

CLUSTERING APPROACH FOR EVENT DETECTION FROM HETEROGENEOUS TWITTER DATA

Ghaemi and Farnaghi (2019) have presented a method that is an improved version of VDBSCAN. This approach is named VDCT, which stands for "Varied Density-based spatial Clustering for Twitter data." The objective of this approach is to recognize and collect geo-tagged events from the data that is available on Twitter, particularly in regions where the data density fluctuates. The VDCT algorithm is able to cluster data of any kind without needing any previous knowledge of the amount of clusters that are included within the data. In addition to being able to distinguish clusters with varying densities, it is particularly successful at maintaining noise control. In light of the fact that our investigation is particularly focused on geographical clustering, we will investigate the benefits and drawbacks of this approach only in respect to spatial clustering.
Positives: The strategy that was presented has been effective in addressing a problem that is encountered by other clustering algorithms. Previous versions of the algorithm used a k-dist network in order to manage data points that had different densities. On the other hand, this technique is only useful for a small number of data points altogether. The fact that this software is able to handle a substantial amount of Twitter data illustrates its high level of efficiency. Instead of using k-distgraph, this makes use of exponential spline interpolation in order to deal with data points that have different densities.

Problems: The input that the user provides for both density-based clustering parameters might have an effect on the quality of the clustering that is produced. .

7 COMPARISON OF VARIOUS ASPECTS OF ALGORITHMS

7.2 DATASETS REVIEW

The DBSCAN method makes use of synthetic sample databases in addition to the benchmark data from SEQUOIA 2000, which is comprised of 62,584 cases and four attributes. O(n2) is the worst-case scenario for its run-time complexity. In its worst-case scenario, the runtime complexity of VDBSCAN is O(n2), and it makes use of a synthetic database that contains data in two dimensions. For the purpose of testing the model, the AGED approach makes use of several multi-dimensional datasets that are routinely used, including Spiral, Wine, Iris, Compound, Zoo, and Canme. The publication, on the other hand, does not provide any information on the difficulty of the program during an actual run. The Geotagged Tweets that are retrieved from the Twitter API are used by DBSTexC. These Tweets include information such as the content, as well as the latitude and longitude coordinates. When considering the worst-case scenario, the run-time complexity of DBSTexC is O(x^2 + xy), where x represents the number of tweets that are relevant to the point of interest and y represents the number of tweets that are irrelevant to the point of interest.

7.3 PERFORMANCE METRICS

Visual Inspection is used by DBSCAN in order to evaluate its model in relation to the CLARANS model that was presented by Raymond T. Ng and Jiawei Han in the year 2002. The performance of VDBSCAN and DBSCAN++ is compared to that of DBSCAN in order to gauge their respective capabilities. A number of different performance

ISSN: 2057-5688

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

measures are used by the AGED model. These metrics include accuracy, silhouette score, Dunn index, entropy, and Pearson gamma. After that, its performance is evaluated in relation to that of DBSCAN. The implementation of ADAPTIVE-DBSCAN and AE-DBSCAN is not evaluated in relation to the implementation of other models. A comparison is made between the F1 score and the DBSCAN score, which is used as the performance metric by DBSTexC. There is a comparison being made between the VDBSCAN algorithm and the VDCT technique.

For the purpose of this inquiry, the datasets and metrics that were employed in the algorithms that were examined are shown in Table 3.

8. CONCLUSION AND FUTURE WORK

This piece provides a demonstration of the capabilities of DBSCAN and analyzes its performance in comparison to three other approaches. VDBSCAN is a method that addresses the issue of variable density, while AGED, AE-DBSCAN, and ADAPTIVE DBSCAN are algorithms that automatically calculate density parameters. The objective of both DBSTexC and VDCT is to identify clusters of high quality by taking into consideration textual information that is also present. DBSCAN++ is an effort to reduce the amount of time that is necessary for calculation. Based on the findings of the study, it is obvious that a substantial amount of research has been undertaken on the subject of calculating ε values. On the other hand, the essential MinPts parameter has not been investigated in any study. In our next endeavors, we will be focusing on the development of a technique that can ascertain both ε and MinPts by taking into account the density of the data. In order to reduce the amount of computing that is required, we want to adapt the DBSCAN++

ISSN: 2057-5688

approach. A detailed assessment of our proposed model will be carried out by comparing it to the first publication via the use of standard clustering criteria. As a result, this will demonstrate that the technique that we have recommended is successful. Research in the future will also investigate the possibility of using the method that was identified to data from social media.

.

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*,
9) *15*(1). https://doi.org/10.1007/s11704-019-9059-3
10) 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
11) 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.
12) 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/ijgi8020

ISSN: 2057-5688

082