ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Analysis of Density Based Spatial Clustering In Data Mining

[1] C.Vinothini, [2] Dr V.Lakshmi Praba
[1] M. Phil., Scholar in Department of Computer Science, Rani Anna Government College, Tirunelveli
[2] Asst Prof, Department of Computer Science, Rani Anna Government College, Tirunelveli

Abstract: **- Data mining involves the association rule learning, classification, summarization, regression, anomaly detection and clustering. Clustering is a data mining technique to group the related data into a cluster and unrelated data into different clusters. Based on the recently described cluster models, there are a lot of clustering that can be applied to a data set in order to partition-ate the information. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is the most well-known density-based clustering algorithm. The aim is to identify dense regions, which can be measured by the number of objects nearest to a given point. Unlike K-Means, DBSCAN does not require the number of clusters as a parameter. It infers the number of clusters on its data, and it can detect clusters of arbitrary shape. Density-based clustering algorithms try to find clusters based on the density of data points in a region. For the experimental work, we have used the milk data set. The results were analyzed and practically tested under MATLAB tools.**

Keywords**: Clustering, Data Mining, Density-based clustering algorithm (DBSCAN) and Spatial domain.**

I. INTRODUCTION

Data mining involves the association rule learning, classification, summarization, regression, anomaly detection and clustering. Clustering is a data mining technique to group the related data into a cluster and unrelated data into different clusters. Based on the recently described cluster models, there are a lot of clustering that can be applied to a data set in order to partition-ate the information. This paper uses the clustering concept, to cluster milk dataset based on its quality. Clustering technique involves several algorithms. In this paper analyzing the algorithm on DBSCAN clustering. Density-based clustering algorithm try to find clusters based on density of data points in a region. This clustering technique uses the density and connectivity that are measured in terms of local distribution of the nearest neighbors. The aim of this paper is to identify the quality of milk suitable for aged persons for the given milk data set with fifty instances.

II. LITERATURE REVIEW

The act of separating meaningful groups of objects that distribute common properties is called clustering and groups having that objects which distribute common properties are called clusters. There are many types of clustering but we will discuss here only two type of clustering which is relevant to our topic i.e. Hierarchical and Partitioned Clustering. Density based methods which is the main concern of our paper belong to Partitioned clustering. [1]. Spatial data mining is the discovery of motivating similarities of individuality and patterns which may exist in large spatial data sets. Spatial clustering is an input perception to get all possible trends and clusters according to given nature of data sets. In the DBSCAN, the density is measured in the form of point which is obtained by including the number of points in a region of specific radius around the point. Points with a certain threshold value and densities form the clusters. Major issue in DBSCAN is the selection of clustering attributes, detection of noise with different densities, and large difference of values of border objects in opposite directions of the same clusters. A point of any object is visited at least once and it may be visited multiple times if it is a candidate of different clusters [2]. The density-based clustering algorithm is a basic data clustering method for finding arbitrary shape clusters as well as for detecting outliers. DBSCAN is compressing data into smaller subsets using k-Means. Then DBSCAN is performed on features of subsets for reducing runtime [3]. Clustering algorithm calculates a density based on the distance metrics that is computed from the data set according to the distance. Then, it selects the points that are dense enough in the space of distance metrics and constructs an abstract space based on these points [4]. Clustering called also unsupervised classification is a process of categorizing a set of data into homogeneous groups (clusters). The elements in each cluster should be similar. Thus, the similarity between individuals in the same cluster (intra-class) must be small and high between the different clusters. This similarity is

considered as a distance measure. In Density-based clustering, the objects are classified based on their regions of density. The density based algorithms have the ability to discover classes of arbitrary shapes and omit noisy objects [5]. For each object in the dataset, the algorithm evaluates the number of neighbors that an object have by counting the number of other objects that are within a proximity radius (minimum R), defined as an input parameter of the algorithm. Based on this calculation, each object is labeled core, border or noise, according to its number of neighbours. If a given object has more neighbours than a minimum value (MinPts), also defined as a parameter of the algorithm, it is classified as core, and all objects reachable from it, either directly (direct neighbours) or indirectly (neighbours of neighbours), are classified as border. All the others object, not reachable from any core, are classified as noise. Each set of objects associated with a core determines a cluster. In density-based clustering that number is derived from the parameters R and MinPts, which also define the density of the clusters to be found [6]. Density based algorithm is not appropriate for data with high variance in density. Generally this algorithm depends upon ordering of degrees within the dataset and it cannot cluster data sets well large distinction in densities. However our projected algorithm for normalized data takes one parameter and deals with high dimensional information set. Additionally free to the starting degrees and able to notice adjacent clusters with varied densities. It also reduces the number of iteration [7].
DBSCAN is a data clustering algorithm. Then, this algorithm proposed by MartinEster, HansPeterKriegel, JorgSander and Xiaoweiu in 1996. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. Density reachability is the initial building block in dbscan. It defines whether two distance close points belong to the same cluster. Points p1 is density reachable from p2 if two conditions are satisfied: (i) the points are close enough to each other: distance (p1, p2) m, where r is a database point. Density connectivity is the last building step of dbscan. Points p0 and pn are density connected, if there is a sequence of density reachable points p1,i2,...,i(n-1) from p0 to pn such that p(i+1) is density reachable from pi. A dbscan cluster is a set of all density connected points [8].

The objective of the data mining technique is to extract information from a large data set and make it into a sensible form for the supplementary purpose. Clustering is a significant task in data analysis and data mining applications. The challenge with clustering analysis is mainly that different clustering techniques give significantly different results on the same data. Moreover, there is no algorithm present which gives all the desired outputs. Because of this, there is extensive research being carried out in „ensembles" of clustering algorithms, i.e. multiple clustering techniques done on a single dataset [9]. This paper presented a survey of Density Based Clustering Algorithm for data mining research that can be helpful for understanding of several clustering algorithms for choosing appropriate algorithm. The type of algorithm that is to be selected depends upon type of clusters that are needed to be finded, type of data set and number of attributes [10].

III.CLUSTERING ALGORITHM AND TECHNIQES

Clustering is to partition data into collection of similar objects. There are many algorithms available for clustering. In this paper, we apply the DBSCAN algorithms using MATLAB tool.

A. DBSCAN ALGORITHM

Density based clustering algorithm has played a very essential role in discovering the non linear shapes structure based on the density. DBSCAN is most widely used density based algorithm. The basic idea of density-based clustering algorithm is that for each instance of a cluster the neighborhood of a radius (Eps) has to include at least a minimum number of instances (MinPts). We have used two important parameters are required for DBSCAN: epsilon ("eps") and minimum points ("MinPts"). The epsilon and minimum point values are 1 and 3. The parameter eps defines the radius of neighborhood around a point x. The parameter MinPts is the least number of neighbors within "eps" radius. For the DBSCAN, the cluster's each data object, who's Eps-Neighbor' objects must smaller than a Minpts. The algorithm defines these data objects as core objects, defines the maximum density of a collection of objects connected as cluster. DBSCAN looks for an object density which start with P about Eps and Minpts from the core object P which never visited from data set D, generate a cluster that contains p and its objects density arriving. The algorithm ends with unvisited objects in the data set D. Thus the visited points were accepted for the clustering under the class defined. As a result every matching neighborhood point clustered under the density region was set as threshold value. Density-based clustering algorithms tries to find clusters based on density of data points in a region.

Algorithmic steps for DBSCAN clustering

Let D = {d1, d2, d3,…,dn} be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

•This process being with an arbitrary starting point that has not been visited.
•Next process is to extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).
•If there are adequate neighborhoods around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).
•If a point is found to be a part of the cluster then its ε (epsilon) neighborhood is also the part of the cluster and the above method from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster are determined.
•Then, a new unvisited point is retrieved and processed, leading to the detection of a further cluster or noise.
•Finally, this process continues until all points are marked as visited.

IV. DATASET DESCRIPTION

The data is collected from various people who are using milk product. In this process, the collected milk product datasets are tested on MATLAB tool using data mining technique. It contains a total of 50 instances. It also includes milk products information and the milk data set takes three attributes as listed below:

•Fat content (range from 1 to 50)
•Cream (range from 1 to 50)
•Butter (range from 1 to 50)

Associated Tasks : Clustering
Number of Instances : 50
Number of Attribute : 3

Sample Data Set

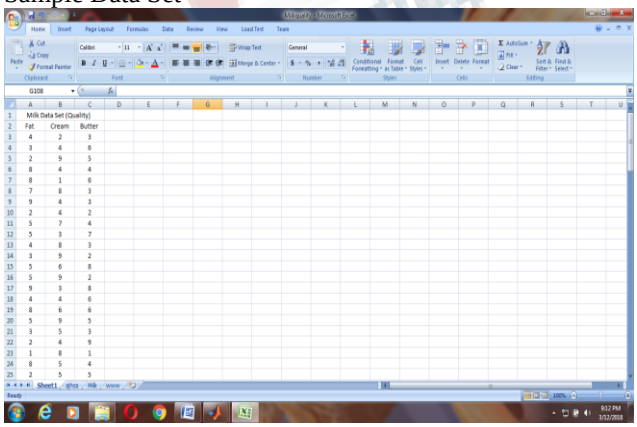

Figure: 1 Sample Data set

V. EXPERIMENTAL RESULT

In this paper, the milk dataset is considered for applying in DBSCAN algorithm. The milk dataset contains the characteristics of various branded of milk and is quality factors obtained from the dataset. The result is analyzed and practically tested in MATLAB. The experiment is carried out on the computer with the configuration such as Pentium (R) Dual-Core CPU T4500 @ 2.30GHz 2.30 GHz, 2GB RAM and Windows 7 Operating System.

The Obtained result is tabulated in Table: 1.

Value	Count	Percent
0 (Noise)	39	78%
1 (Cluster 1)	4	8%
2 (Cluster 2)	3	6%
3 (Cluster 3)	4	8%

The quality of milk is tested based on the following three attributes

•Fat content
•Cream
•Butter

The milk quality is suggested by doctors as suitable based on the following condition for aged people

Fat content < 5 and Cream < 5 and Butter < 5

The implemented result shows that 78% is suitable for them and 22% is not suitable. Among the 50 instances, 4 of Fat content, 3 of Cream and 4 of Butter are not suitable. This is illustrated in Figure: 2

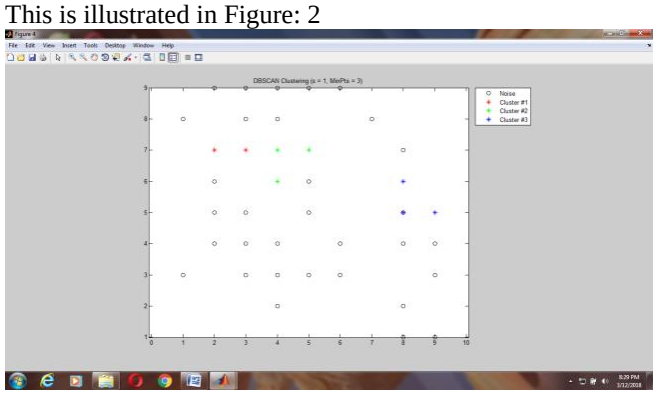

Figure: 2 Result of the DBSCAN Algorithm

VI. CONCLUSION

The aim of this paper is to identify the quality of milk suitable for aged persons for the given milk data set with fifty instances. In this process the clustering of the milk

data set is done with three attributes namely fat, cream and butter content using DBSCAN algorithm. The DBSCAN algorithm process is done on the milk data set in MATLAB2014a for the clustering of quality factors. The obtained result from these attributes yielded 78% quality of milk suitable for aged persons.

REFERENCES

[1] Pooja Batra Nagpal and Priyanka Ahlawat Mann," Comparative Study of Density based Clustering Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011.

[2] Arvind Sharma,1 R. K. Gupta,2 and Akhilesh Tiwari," Improved Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data", Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2016, Article ID 1564516, 9 pages.

[3] Son T. Mai,Ira Assent and Martin Storgaard," An Efficient Anytime Density-based Clustering Algorithm for Very Large Complex Datasets", KDD '16, August 13-17, 2016, San Francisco, CA, USA.

[4] Mariam Rehman and Syed Atif Mehdi, "Comparison of Density-Based Clustering Algorithms", www.researchgate.net/publication/242219043.

[5] Hajar Rehioui, Abdellah IDRISSI, Manar ABOUREZQ and Faouzia ZEGRARI, "DENCLUE-IM: A New Approach for Big Data Clustering", The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016), Procedia Computer Science 83 (2016) 560 – 567.

[6] Guilherme Andrade, Gabriel Ramos, Daniel Madeira,Rafael Sachetto, Renato Ferreira and Leonardo Rocha," G-DBSCAN: A GPU Accelerated Algorithm for Density-based Clustering", International Conference on Computational Science, ICCS 2013", Procedia Computer Science 18 (2013) 369 – 378.

[7] Nidhia and Km Archana Patel, "An Efficient And Scalable Density-Based Clustering Algorithm For Normalize Data", 2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016)", Procedia Computer Science 92 (2016) 136 – 141.

[8] B.L. Krishna, P.Jhansi Lakshmi and P.Satya Prakash," Combination of Density Based and Partition Based Clustering Algorithm-DBK Means", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012,4491 – 4494.

[9] K. Chitra1 , Dr. D.Maheswari2," A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 6, Issue. 8, August 2017, pg.109 – 115.

[10] Rashi Chauhan, Pooja Batra and Sarika Chaudhary, "A Survey of Density Based Clustering Algorithms", International Journal of computer science Trends and Technology (IJCST) Vol. 5, Issue 2, April - June 2014.