Solid State Technology
Volume: 63 Issue: 5
Publication Year: 2020

Comparative Study of Density-based Clustering Algorithms

Bhumika S. Arora, Research Scholar[1], Dr.Vijay Chavda[2], Dr Bhadresh R. Pandya[3]

[1]*Ph.D. Scholar, Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat, India*

[1]bhumika.mca@gmail.com

[2]*Principal, N.P College of Computer Studies and Management, Kadi, Gujarat, India*

[2]dr.vijaychavda@gmail.com

[3]*Head of Department, M.Sc.IT & B.Sc.CS, Gandhinagar, Gujarat, India*

[3]prof_brpandya@yahoo.in

Abstract: Data mining and exploratory data analysis deal with large datasets wherein one of the objective is to group objects such that objects falling under the same group have more similarities than the objects of two different groups, these groups are called clusters. Here similarity of objects is measured in terms of distance between the objects. The process of grouping objects is called clustering. There are numerous applications of clustering in variety of fields and data involved in each varies a lot. This has led researchers to come out with various methods and models of clustering. There are different types of clustering methods: Hierarchical, Partitioning, Density Based, Grid based and Graph Based. This paper presents a study of various methods of clustering and discusses briefly on various Density based clustering methods.

Keywords: *Clustering, Process of Clustering, Types of Clustering, Clustering methods, Cluster analysis.*

1. Introduction:

Clustering is the process that prepares different groups of objects which are called clusters where objects in the same group have more similarity than the objects in the different groups.

There are numerous methods of cluster analysis available which makes it a difficult task for user in the situation when clustering is required.

Concept of clustering and its applications:

Clustering is applied in many different kind of fields.

- New patterns are being looked for the in the exploratory data analysis without any specific interpretation which may raise new research questions.
- Demarcation of species in animals, birds, plants
- Geospatial analysis
- Classification of diseases
- Efficient data organization
- Image segmentation

Clustering is a process where similar objects are put in one cluster and dissimilar objects are put into another cluster. The process or method differs on case to case basis and there is no universally accepted method of clustering and defining a cluster. Meaning of a "cluster" may differ a lot based on the application or the field it is being used. Depending on the area of application an appropriate clustering method is adapted which in turn defines cluster.

2. Types of Clustering Methods

Commonly used methods for Clustering are Hierarchical, Partitioning, Density Based, Grid based and Graph based.

2.1 Hierarchical Method

In the hierarchical clustering method [1] clusters are build as hierarchical decomposition based on a criterion. This can be visualized as a tree like diagram which records the sequence of merges or splits, known as dendrogram. This approach allows exploration of data with different levels of granularity. Clusters are obtained by 'cutting' the dendrogram at the different levels.

Method of hierarchical clustering is further categorized into bottom-up and top-down approach. In bottom-up approach, each object is treated as a single cluster. These clusters are then merged recursively with most similar clusters until all clusters are merged into one big cluster. Where as in top-town approach, all objects are put into a single large cluster at the beginning. Iteratively, each cluster is divided into two most heterogeneous clusters. These iterations continue till all objects are into their own individual cluster.

Algorithms that follow hierarchical clustering methods are LEGCLUST, BRICH, CURE, Chemeleon, ROCK, Linkage, Leaders–Subleaders, Bisecting k-Means and AGNES.

2.2 Partitioning Method

Unlike hierarchical clustering [9] which generated successive level of clusters by iterative merging or splitting; partitioning method of clustering assigns a set of objects into K clusters with no hierarchical structure. Algorithms under this method, work to minimize a certain clustering criterion by relocating data points between different clusters iteratively, till optimal partition is attained. A database D of N objects is partitioned into a set of K clusters; here the number of clusters K is predetermined. One of the most widely used criteria is the sum of squared error function. Some of the algorithms under this method are K-means, K-mediods (PAM), CLARA and CLARANS.

2.3 Density based Methods

Density based clustering algorithms [15] group objects based on the density of regions. These algorithms identify low-density and high-density regions separately, which allows them to find clusters having different shapes. Since they work based on density, they are capable to handle noise. Density-based clustering algorithms can automatically detect the clusters and do not require to specify the parameter for number of clusters. Clusters formed based on density do not limit themselves to any shape and are easy to understand.

2.3.1 DBSCAN (Density based Spatial Clustering of Application with Noise)

DBSCAN algorithm [16][17] works based on two parameters viz. 1) Eps: Specifies radius of neighborhood around data point p and 2) minPts: Specifies number

of minimum data points in the neighborhood to identify it as a cluster. Using these two values, the algorithm discovers clusters by using concept of density reachability and connectivity. Since density reachability is non-linear, this algorithm can discover clusters with different shapes. All data points of the data set are categorized into i) Core points ii) Border points, and iii) Noise points.

DBSCAN algorithm has got a limitation that it is can't detect clusters having different densities.

2.3.2 VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise)

VDBSCAN [18] arrives at the value of parameter Eps and MinPts by finding distance of k^{th} nearest neighbor of the point, after which it finds sharp change in the distance, which is called k-dist. This allows it to come out with different partitions of the given data sets and with multiple Eps values. Using which it generates multiple clusters with different densities.

The challenge here is that the magnitude of impact for finding k-dist depends fully on the characteristic of the data set.

2.3.3 DVBSCAN (Density Based Algorithm for discovering Density Varied Clusters in Large Spatial Databases)

DVBSCAN – algorithm [19] works by calculating growing cluster density mean and cluster density variance of the core point. It allows expansion of the neighborhood of core point when variation of cluster density is less than or equal to the limit set. This algorithm can find clusters which have relatively uniform density even though they are not separated by sparse regions.

The algorithm can be improved on performance and automatics calculation of parameters.

2.3.4 DBCLASD (Distribution- Based Clustering Algorithm for Mining Large Spatial Databases)

The DBCLASD algorithm is [20] works efficiently on large spatial databases and doesn't need any input parameter. This is an incremental algorithm which assigns points to initial cluster if its nearest neighbor distance is as per the expected distance distribution. In this incremental approach, candidate points are not discarded but they are considered later for the assignment to suitable cluster. The algorithm assumes that there is uniform distribution of points within the cluster.

2.3.5 ST-DBSCAN (Spatial Temporal Density Based Clustering)

The ST-DBSCAN algorithm is [21] an improved version of DBSCAN which can discover clusters based on spatial, non-spatial and temporal attribute values of the objects. It has three modifications, first it can discover cluster on spatial-temporal data, second it has introduced concept of density factor which assigns degree of density to the cluster and third it provides comparison of average value of a cluster to the new coming value.

The algorithm can be improved to automatically generate input parameter and also by improving performance of it.

2.3.6 GDBSCAN (Generalized DBSCAN)

The GDBSCAN algorithm [22][23] includes spatial and nonspatial attributes of the objects while clustering. It discovers density-connected sets and noise in spatial database. The algorithm works based on three parameters, 1) neighborhood predicateNPred, 2) minimum weightMinCard and 3) a weight function wCard.

Enhancement in this work may include heuristic function to determine the parameter wCard which is not based on cardinality function.

2.3.7 DENCLUE (DENSITY BASED CLUSTERING)

DENCLUE algorithm [16][24] works based on the Kernel Density Estimation technique which is aimed to find dense regions. It was mainly developed to classify large multimedia databases which have high-dimensional data and contain large amount of noise. The algorithm has two step process, in the first step, it constructs a hyper-rectangle (map) of the database and in the second step, it identifies clusters from highly populated cubes.

DENCLUE algorithm needs to be improved for more efficiency or for larger datasets.

2.3.8 OPTICS (Ordering Points to Identify the Clustering Structure)

The OPTICS algorithm [25] uses random values for Eps which are used to identify clusters by generalizing technique of DBSCAN. This algorithm find clusters with varying density. It calculates outlier score for each point considering distance from its closest point.

Future research on this can be based on improving efficiency to support hyper sphere range queries for high-dimensional spaces having no index structures.

2.3.9 CLIQUE (CLustering In QUEst)

CLIQUE algorithm [1][22][23] is useful for clustering high dimensional data. It combines techniques of density based clustering and grid based clustering. The algorithm uses two parameters, 1) xi – number of intervals and 2) tau – density threshold. It divides n-dimensional data space into several non-overlapping rectangular units and identifies dense regions among those. A cluster here is a maximal set of connected dense units. It also uses a-priori algorithm to identify connected dense units.

2.3.10 BRIDGE

The BRIDGE algorithm [27] combines techniques of K-means and DBSCAN algorithms to overcome limitations of each other. It enables DBSCAN to handle very large data whereas it also removes noisy points by improving technique of K-means. It performs K-means first followed by density based clustering. It helps setting density threshold parameter properly. This approach makes it faster and computationally cost effective.

2.3.11 LSDBC (Locally Scaled Density Based Clustering)

The LSDBC algorithm [28] works on the technique of local scaling, has got two input parameters, 1) k – used to order points according to their distance to their kth neighbor and 2) α – used to determines the boundary of the current cluster expansion based on its density. The local scaling technique separate clusters using local

statistics of the points. These parameters help to know how dense the region is around each point. Beginning with higher density regions, it connects points of dense regions until the density fall below the threshold.

2.3.12 CUDA-DClust

The CUDA-DClust algorithm [29] is a density based clustering algorithm which uses high computational power of GPU under NVIDIA's CUDA architecture and programming model. It produces multiple clusters assigning a single point to simultaneously to the GPU. The algorithm can perform multiple distance operations between the objects in each thread simultaneously.

The limitation of this approach is that the collision matrix is stored in off-chip device memory of GPU, which is very high cost of memory access.

2.3.13 UDBSCAN

The UDBSCAN algorithm [30] is designed to work on uncertain objects. Uncertain objects are the one which have certain attributes whose precise value can't be defined. This may be due to various reasons including data acquisition or property of the object itself. The U-DBSCAN algorithm uses a deviation function that approximates value of such attribute and creates clusters using an associated probability density function.

2.3.14 Grid based DBSCAN

The Grid based DBSCAN algorithm [31] uses the grid to reduce amount of data processed by DBSCAN and improves efficiency. In the first step, the data points are assigned to grids based the values of their attributes. Noise points are removed using grid method. After that, grid data center is used as input to DBSCAN which reduces data processing time.

2.3.15 GRP-DBSCAN (GRP-DBSCAN Algorithm with Referential Parameters)

The GRP-DBSCAN algorithm [32] with Referential Parameters chooses Eps and MinPts parameters automatically. It combines grid partition technique and mutli-density based clustering algorithm. It can discover clusters with any arbitrary shapes and can identify noise points.

This technique has limitation of handling datasets having different densities.

2.3.16 DBCURE

DBCURE algorithm [33] parallelizes operations using MapReduce technique and can find clusters having different densities. DBCURE uses concept of ellipsoidal τ-neighborhood. This two-step approach of DBCURE, it selects an unvisited core point to start with and inserts that to set S. In the next step, it retrieves all points that are density-reachable from the set S, inserting its ellipsoidal τ-neighborhood. Improved version of this algorithm it, DBCURE-MR can find several clusters in parallel.

2.3.17 Fast-DBSCAN

This algorithm [34] employs Groups method to accelerate neighbor search. The Groups method partitions dataset into graph where vertex represent a group and an edge between two vertices is drawn if the two groups are reachable. Groups method builds graph based index on the data, which scales well for high dimensional

datasets. This method can handle large amount of noise in the data efficiently and produces results identical to that of DBSCAN with better performance.

2.3.18 DENDIS

DENDIS is a hybrid sampling algorithm [35] which uses density-based concept while managing distance to ensure space coverage. It combines concept of DENsity and DIStance when selecting observations for the sample from the large datasets. It chooses a new item for the sample which is farthest from the representation in the most important group.

This algorithm can be improved by making it self-tuning, capable of finding by itself the suitable granularity to reach a given level of accuracy.

2.4 Grid-based methods

In this method [1] data space is divided into number of cells known as grids. Clustering is performed on these grids which may contain several data points, which makes the process faster as it processes grids rather than all the data points. Complexity and processing time for the clustering depends on populated grid cells. Examples of this algorithm are STING, CLIQUE, Wave Cluster, OptiGrid, GDILC and MAFIA

2.5Graph-based methods

Graph-based clustering [1] are of two types, 1) between graph – prepares clusters among set of graphs, 2) Within graph – prepares clusters of nodes/edges within a single graph. Example of graph-based algorithms are Shared Nearest Neighbour, Between-ness Centrality Based, Highly Connected Components, Maximum Clique Enumeration, Kernel K- means algorithm.

3. Comparison of Density based Clustering Algorithm

Cluster Algorithm	Complexity	Input Parameter	Capability of tackling High/Large Dimensional Data	Sensitivity to outliers	Identifies Arbitrary Cluster Shapes
DBSCAN	$O(n^2)$ Using R-tree index- O(n * log n)	Eps, Minpts	Yes	Yes	Yes

VDBSCAN	O(n)	Automatically generated	Yes	Yes	Yes
DVDBSCAN	$O(n^2)$	Radius , 2 threshold parameters	Yes	Yes	Yes
ST-DBSCAN	O(n * log n)	Eps1 (Distance parameter of spatial attribute), Eps2 (Distance Parameter of non-spatial attribute), Minpts and $\Delta\in$	Yes	Yes	Yes
DBCLASD	$O(3n^2)$	None	Yes	Yes	Yes
GDBSCAN	O(n * log n)	Eps, Minpts	Yes	Yes	Yes
DENCLUE	O(log| n|)	Two Parameters σ, ξ	Yes	Yes	Yes
OPTICS	Without index- $O(n^2)$ With Spatial index- O(n log n)	Eps, Minpts	Yes	Yes	Yes
CLIQUE	O(n)	Size of the Grid , Minimum Number of Points within each Cell	Yes	Yes	Yes
BRIDGE	O(n log n)	Dataset D, and K(Number of cluster)	Yes	Yes	Yes
LSDBC	O(n log n)	K and α	Yes	Yes	Yes

CUDA-DClust	O(n * log n)	Eps, Minpts	Yes	Yes	Yes
UDBSCAN	> $O(n^2)$	Uncertain objects	Yes	Yes	Yes
Grid-based-DBSCAN	O(n * log n)	Grid size, noise pts	Yes	Yes	Yes
GRP-DBSCAN	n^2	Eps,MinPts	Yes	Yes	Yes
FAST-DBSCAN	O(n + nd)	5*Eps	Yes	Yes	Yes
DENDIS	O(n * s)	Xi,I	Yes	Yes	Yes

4. Conclusion

The objective of data mining and exploratory analysis is to transform large amount of data into an understandable form which can be useful. Clustering is one of the important task to identify group of objects which are more similar within the group than the objects of two different groups. There are different methods of clustering including hierarchical-based, partitioning-based, density-based, grid-based and graph-based. Density-based clustering methods are intended to form clusters based on density of regions. Meaning and method of "cluster" differs depending on the application it is being used. One of the main are where density-based methods of clustering are used is of spatial databases. Many density-based algorithms, as briefly discussed in the paper, have been developed to address different applications and improvement of certain features of the base algorithms. With ever increasing complexity and scale of spatial databases along with non-spatial attributes and variety of application areas, there is always a scope of new algorithms to come out with different types of "clusters".

5. References

[1] K.Kameshwaran,& K.Malarvizhi2 (2014). Survey on Clustering Techniques in Data Mining, (IJCSIT)International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2272-2276
[2] Jorge M. Santos, Joaquim Marques de Sa´, &Luı´s A. Alexandre (2008). LEGClust: A Clustering Algorithm Based on Layered Entropic Subgraphs, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 30.
[3] Yogita Rani & Harish Rohil(2013). A Study of Hierarchical Clustering Algorithm , International Journal of Information and Computation Technology, ISSN 0974-2239 Volume 3, 1115-1122
[4] Zhang, T., Ramakrishnan, R. & Livny. M., (1997). BIRCH: A New Data Clustering Algorithm and Its Applications. Data Mining and Knowledge Discovery 1, ACM SIGMOD, 141–182
[5] Karypis, G., Eui-Hong Han, & Kumar, V. (1999).Chameleon: hierarchical clustering using dynamic modeling. Computer IEEE, 32(8), 68–75.
[6] SUDIPTO GUHA, RAJEEV RASTOGI, & KYUSEOK SHIMS (2000). ROCK: A ROBUST

CLUSTERING ALGORITHM FOR CATEGORIAL ATTRIBUTES, Published by Elsevier Science Ltd. Information Systems Vol. 25, No. 5, pp. 345-366.
[7] R. Sibson (1973). SLINK: An optimally efficient algorithm for the single-link cluster method, The Computer Journal, Volume 16, Issue 1, 1973, Pages 30–34.
[8] Kaufman, L., & Rousseeuw, P. J. (1991). Finding Groups in Data: An Introduction to Cluster Analysis.Journal of the American Statistical Association,Vol. 86, No. 415 (Sep., 1991), pp. 830-832.
[9] Rui Xu, & Donald Wunsc (2005). Survey of Clustering Algorithms, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3.
[10] Swarndeep Saket J, Dr. Sharnil Pandya (2016). An Overview of Partitioning Algorithms in Clustering Techniques International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 6.
[11] Madhu Yedla, Srinivasa Rao Pathakota & T M Srinivasa (2010). Enhancing K-means Clustering Algorithm with Improved Initial Center, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 121-125.
[12] A. P. REYNOLDSj, G. RICHARDS, B. DE LA IGLESIA & V. J. RAYWARD-SMITH (Springer 2006). Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms, Journal of Mathematical Modelling and Algorithms 5: 475–504.
[13] Raymond T. Ng & Jiawei Han (2002), CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 14, NO. 5.
[14] Tülin Inkayaa, Sinan Kayalıgil b, Nur Evin Özdemirel (2014). An adaptive neighbourhood construction algorithm based on density and connectivity,Elsevier.
[15] Aina Musdholifah & Siti Zaiton Mohd Hashim(2013). Cluster Analysis on High-Dimensional Data: A Comparison of Density-based Clustering Algorithms, Australian Journal of Basic and Applied Sciences, 7(2): 380-389.
[16] Martin Ester, Hans-Peter Kriegel, Jiirg Sander & Xiaowei Xu(1996).A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, KDD-96 Proceedings.
[17] Peng Liu, Dong Zhou & Naijun Wu(2007).VDBSCAN:Varied Density Based Spatial Clustering of Applications with Noise, IEEE, International Conference on Service Systems and Service Management.
[18] Anant Ram, Sunita Jalal, Anand S. Jalal & Manoj Kumar(2010). A Density based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases. International Journal of Computer Applications (0975 – 8887) Volume 3 – No.6.
[19] Xiaowei Xu, Martin Ester, Hans-Peter Kriegel & Jörg Sander(1998). A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases, IEEE, Proceedings 14th International Conference on Data Engineering.
[20] Derya Birant & Alp Kut(2006).ST-DBSCAN: An algorithm for clustering spatial–temporal data, Elsevier,Data & Knowledge Engineering 60 ,208–221.
[21] Jörg Sander, Martin Ester, Hans-Peter Kriegel & Xiaowei Xu (1998).Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications,Springer,Data Mining and Knowledge Discovery volume 2, pages169–194.
[22] Pradeep Singh & Prateek A. Meshram(2017). Survey of Density Based Clustering Algorithms and its Variants, IEEE Proceedings of the International Conference on Inventive Computing and Informatics.
[23] Hajar REHIOUI∗, Abdellah IDRISSI, Manar ABOUREZQ & Faouzia ZEGRARI (2016). DENCLUE-IM:A New Approach for Big Data Clustering, Elsevier The 7th International Conference on Ambient Systems, Networks and Technologies , Procedia Computer Science 83 (2016) 560 – 567.

[24] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel & Jörg Sander(1999). OPTICS: Ordering Points To Identify the Clustering Structure,Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA.
[25] Ms K.Santhisree & Dr. A Damodaram(2011).CLIQUE: Clustering based on density on Web usage data: Experiments and Test results,IEEE 3rd International Conference on Electronics Computer Technology.
[26] Manoranjan Dash, Huan Liu & Xiaowei Xu(2001). 1 + 1 > 2': Merging Distance and Density Based clustering, IEEE Proceedings Seventh International Conference on Database Systems for Advanced Applications.
[27] Ergun Bic¸ici & Deniz Yuret(2007). Locally Scaled Density Based Clustering, Springer-Verlag Berlin Heidelberg, ICANNGA, Part I, LNCS 4431, pp. 739–748.
[28] Christian Böhm, Robert Noll, Claudia Plant & Bianca Wackersreuther (2009). Density-based clustering using graphics processors, Proceedings of the 18th ACM conference on Information and knowledge management Pages 661–670.
[29] Apinya Tepwankul & Songrit Maneewongwattana(2010). U-DBSCAN: A density-based clustering algorithm for uncertain objects, IEEE 26th International Conference on Data Engineering Workshops.
[30] Li Ma, Lei Gu, Bo Li , Sou yi Qiao & Jin Wang(2014). G-DBSCAN: An Improved DBSCAN Clustering Method Based On Grid, Advanced Science and Technology Letters Vol.74 ,pp.23-28.
[31] Huang Darong & Wang Peng(2012). Grid-based DBSCAN Algorithm with Referential Parameters, Elsevier,International Conference on Applied Physics and Industrial Engineering, pp : 1166-1170.
[32] Younghoon Kim, Kyuseok Shim, Min-Soeng Kim & June Sup Lee(2013). DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce, Published by Elsevier Ltd.
[33] K. Mahesh Kumar & Dr. A. Rama Mohan Reddy (2016). A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method.
[34]FrédéricRos, Serge Guillaume(2016).DENDIS: A new density-based sampling for clustering algorithm, Published by Elsevier Ltd