International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 5th Jan 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 01

DOI: 10.48047/IJIEMR/V12/ISSUE 01/03

Title **A Survey on Performance of Various Density Based Clustering Techniques on Stream Data and Its Applications**

Volume 12, ISSUE 01, Pages: 33-42

Paper Authors
Mallesh Babu S, Dr.V.K. Sharma

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A Survey on Performance of Various Density Based Clustering Techniques on Stream Data and Its Applications

Mallesh Babu S[1], Dr.V.K. Sharma[2]
[1]Research scholar, Department of Computer Science and Engineering, Bhagwant University, Ajmer, malleshbabu41@gmail.com
[2]Professor, Dept. of EEE, Bhagwant University Ajmer, viren_krec@yahoo.com

Abstract
Clustering of stream data is a one of the challenging task in data mining. Stream data is a sequence of ordered data without any boundaries. Now a days, stream data is generated by multiple real time applications like user clicks on WWW, smart phone GPS tracking data, data from sensor devices, stock market data, communication network data, patients continuous monitoring data, and so on. In general, clustering of stream data is performed in two steps, one is data summarization into micro clusters, and second one is merging of micro clusters into clusters based on similarities. In this paper, we conducted details survey on cluster data mining techniques, algorithms that are used to cluster data streams, number of techniques used to cluster stream data using density based clustering, and various methods used to form grid clusters, and its applications. In this study, we compared various density based clustering, types of data used, objectives, input data applied, outcome generated, time complexity of clustering methods, and space complexity of each clustering methods.

Keywords: *Data Mining, Clustering, Density Based Clustering, Complexities, DBSCAN, Applications*

1. Introduction

KDD is a process of identifying useful patterns from data warehouse. In spatial data mining (SPM) use spatial data with multi dimensions. In spatial database store data which is obtained from satellite, patient x-ray data, and so on [23]. SPM is a method of identifying patterns from spatial database and process is hard when compared with conventional clustering methods. In SPM, handle data types, data models, relationships, topology, preprocessing, and efficiency. Spatial data can be applied to decision tree

classification, neural networks, fuzzy logic, many classification and clustering algorithms [22]. Day by day development in networking, users may use many network devices like smart phone, sensor devices to capture data, hours together use of social media, these may generate large volume of data. With human understanding capability it is difficult to analyze all these data, number of clustering techniques are introduces to form clusters [24].

Clustering is used in many applications like medical field, environment monitoring, outlier detection, recommendations, and so on [19]. Clustering is a one of the data mining technique useful to arrange input data into smaller groups called clusters. Clustering algorithms broadly divided into seven categories, partitioning, grid, hierarchical, density, model, graph, and combination of these techniques [20]. In density based techniques user need not to mention number of clusters and algorithm accept input of any size and shape. In this paper we study the complete set of algorithms used for clustering stream data, its applications, types of data used, and evaluated performance of each technique.

II. Related Work

Clustering is a method of grouping data into smaller clusters based on distance measures like similarity or dissimilarity between data samples [25]. DBSCAN is a method to cluster data samples based on density [17]. DBSCAN has more advantage and will be able to calculate two parameter values EPS and MINPTS with a challenge of highly dependent data samples [26]. To solve this problem authors proposed hierarchical based clustering (DBHC). They used KNN to calculate values of EPS and MINPTS [21]. In general real time database data is not uniformly distributed and because of this it will generate more EPS values [16]. Based on number of EPS values DBHC method will run DBSCAN many times to produce clusters [18]. DBHC produce less number of clusters by merging same samples when compared with DBSCAN method [13]. Authors used UCI repository data as input to DBHC, execute algorithm to produce clusters, and evaluate performance with other clustering algorithms [1].

Authors [2] worked on arbitrary data with distance parameter and apply density based clustering (DBC) algorithm to produce clusters with high quality [7]. To extract knowledge patterns they used

similarity measures [15]. Authors designed experimental framework to accept high dimensional data as well as big size data and used approximation technique to form clusters [12]. Performance of their algorithm is compared with conventional K-means and proposed method of clustering is shows better results than k-means algorithm [2]. Clustering technique is used in many applications [9]. DBSCAN is one of the unsupervised techniques to form high density clusters by separating samples with low density [6]. Authors proposed a new method AOAOBL based on optimization learning technique to overcome problems of DBSCAN [11]]. They evaluate proposed algorithm with ten types of datasets and their method shows better result than DBSCAN [3].

DBSCAN has many features then other clustering algorithm, number of clusters need not to be fixed in advance, algorithm will apply for any type of dataset with various shapes, outliers are identified during clustering process, and there is no outlier sensitivity maintained [10]. Disadvantages of DBSCAN are required more computations to calculate values when dense values varied during clustering process [8]. Authors [4] proposed two phase method of clustering, first step is improved version of DBSCAN, and second step is DP (Density peak) algorithm working based on decision tree at the time of choosing centers of clusters [14]. Cluster centers are selected automatically based on decision tree, and initial centers of clusters changed automatically [5].

III Comparative Study

$n_p = number\ of\ points,$
$n_c = number\ of\ clusters,$
$n_{cs} = number\ of\ constraints,$
$n_A = number\ of\ Agents,$
$T_D = Time\ required\ to\ Process,$
$D_s = Data\ size,$
$n_{pc} = number\ of\ potential\ clusters,$
$n_{ipc} = number\ of\ interested\ potential\ clusters,$
$L_C = level\ of\ clustering,$
$n_{dpc} = number\ of\ deleted\ potential\ clusters,$
$n_{cg} = number\ of\ cluster\ grids,$
$n_g = number\ of\ grids,$
$T_O = Order\ of\ Tree,$
$S_j = size\ of\ job,$
$IDS - Intrusion\ Detection\ System,$
$ME = Monitoring\ Environment,$
$S_{dg} = size\ of\ grid.$

More number of conditions is used in semi supervised technique of clustering. All

previous researchers worked on either condition based or parameter based clustering methods. In conditional based clustering method, conditions are applied to divide the data samples into smaller group of samples called clusters, and whereas parameter based clustering use certain measures and divide data samples into clusters by following supervised learning method. Almost all types of clustering follow conditions based clustering methods to cluster data samples, if conditions are good algorithm cluster data samples accurately, and if conditions are not correctly fixed then algorithm shows low performance and accuracy also low. We conduct detailed survey on density based clustering algorithms, types of data used in each algorithm, list of input parameters, and list of outcomes, and complete data is listed in table 1.

S.No	Algorithm	Kind of Data	Objective	Input	Output
1	DENS	continuous data	clustering streams	radius, threshold, weight, decay	arbitrary type of clusters
2	CDENS	continuous data	constraint based clustering	cradius, oradius, min neighbors, constraints, decay	constraint based arbitrary type of clusters
3	RDENS	continuous data	accuracy of clusters	radius, threshold, weight, decay	arbitrary type of clusters
4	HDENS	continuous data	quality of clusters	radius, threshold, weight, decay	arbitrary type of clusters
5	SDS	continuous data	clustering within sliding window	radius, wsize, weight	clusters within wsize
6	SOS	continuous data	automation of clustering	radius	threshold
7	FLOCKS	continuous data	flocks based clustering	radius, threshold, weight, decay	arbitrary type of clusters
8	SOPTICS	continuous data	visualization of clusters	m_list, distance	clusters within time
9	HDDS	categorical data	high dimensional data clustering	radius, threshold, weight, decay	arbitrary type of clusters
10	PDCS	continuous data	high dimensional data clustering	radius, threshold, weight, decay	arbitrary type of clusters
11	ExCC	continuous data	clustering of heterogeneous data	grid values of granularities	arbitrary type of clusters
12	DCUS	continuous data	clustering of uncertain values	dimension, density	arbitrary type of clusters
13	MRS	continuous	accuracy of	data, threshold,	clusters with

		data	clusters	decay	resolution
14	DDS	continuous data	quality of clusters	data, threshold, decay	arbitrary type of clusters
15	DUCS	not defined	only one scan for clustering	data in the form of streams	connected components of cluster
16	PKSS	continuous data	high dimensional data clustering	tree, density	arbitrary type of clusters
17	DENGRISS	continuous data	clustering within window	data, window size	arbitrary type of clusters

Table 1: Density based algorithms data, objective, input, and output

Active learning algorithm is also one of the clustering algorithm where as a pair of conditions are used to form clusters. In this method of clustering authors [5] identify K-Neighbors of various clusters and then select data points and conditions. Min-Max parameter is also used to choose data samples of high uncertain values and this method is not appropriate to cluster data samples with high dimensions. Active learning algorithm generates conditions based on KNN graph and these conditions may degrade performance of a clustering technique. We conduct detailed survey on density based clustering algorithms, types of algorithm is used to cluster data samples, type of data used like noisy data, high dimensional data, and evolving data, and complete data is listed in table 2.

S.No	Algorithm	Evolving Data	Noisy Data	High Dimensional data
1	DENS	yes	yes	no
2	CDENS	yes	yes	no
3	RDENS	yes	yes	no
4	HDENS	yes	yes	no
5	SDS	yes	yes	no
6	SOS	yes	yes	no
7	FLOCKS	yes	yes	no
8	SOPTICS	yes	yes	no
9	HDDS	yes	yes	yes
10	PDCS	yes	yes	yes
11	ExCC	yes	yes	no
12	DCUS	no	yes	no
13	MRS	yes	yes	no
14	DDS	yes	yes	no
15	DUCS	yes	yes	no
16	PKSS	yes	yes	no
17	DENGRISS	yes	yes	no

Table 2: Density based algorithms and type of data used in clustering

Earlier methods on clustering, DBSCAN, and improved versions of these methods are grouped data samples based density. In DBSCAN first calculate density of each data samples by identifying points and then applied threshold value to identify inside data points, boundary data points, and outliers. All inside data points are grouped into cluster first, boundary points assigned to clusters, and outliers are removed from data points. DENCLUE

clustering algorithm is working on density of kernels, and or neighbors of nearest. We conduct detailed survey on density based clustering algorithms, applications, parameters used for evolution, time and space complexities of clustering algorithm, and complete data is listed in table 3.

S.No	Algorithm	Application	Metrics	Time complexity	Space
1	DENS	IDS	Purity	$O(n_c)$	n_c
2	CDENS	ME	RI	$O(n_c + n_{cs})$	$n_c + n_{cs}$
3	RDENS	IDS	Purity	$O(n_c + T_D)$	$n_c + D_s$
4	HDENS	IDS	Purity	$O(n_c)$	n_c
5	SDS	IDS	Purity	NA	L_W
6	SOS	IDS	Purity	$O({n_p}^2 log_n)$	n_c
7	FLOCKS	IDS	Purity	$O(n_c) + O(n_a)$	$n_c + n_a$
8	SOPTICS	ME	NA	n_c	$O(n_c * \log(n_c))$
9	HDDS	ME, IDS	Purity	$O(n_c) + O(n_{pc})$	n_c
10	PDCS	IDS	Purity	$O(n_c) + O(n_{ipc}) + O(n_{dpc})$	n_c
11	ExCC	IDS	Purity	$O(n_{cg})$	$n_g + s_{dg} + s_g$
12	DCUS	ME	Quality	$O(n_g)$	n_g
13	MRS	IDS	Purity	$O(n_g * L_c) + O(2^{n_g} * L_c) + O(n_g * \log(n)$	$n_g * L_c$
14	DDS	IDS	NA	$O(n_g^2)$	n_g
15	DUCS	ME	Quality	$O(n_g)$	n_g
16	PKSS	IDS	Purity	$O(\log T_o), O(T_o)$	$log_{T_o}^{n_g}$
17	DENGRISS	NA	Purity	$O(n_g)$	n_g

Table 3: Density based algorithms Applications, metrics, time complexity, space

DBSCAN and DENCLUE clustering algorithms are not accepting object stream data to cluster. A stream data is a sequence of ordered data samples without boundaries. All data cannot be stored into stream and also random access of all data is not feasible. Streams are used to handle dynamic data and clusters are created and some old clusters are disappearing. Over last decade, researchers on this problem proposed number of algorithms, all are working in two phases, and also working in dual mode (offline or online). In order to solve this problems micro clustering is introduced, where all data points' density values are calculated and form micro clusters, and again this micro clusters are clustered into main clusters. We conduct

detailed survey on density based clustering algorithms, execution time of clustering algorithms, and complete data is tabulated in table 4 and 5.

S.No	Algorithm	Method	Time Complexity
1	DBSCAN	Density	$O(n^2)$
2	DBSCAN	R* Tree	$O(n\ lon\ n)$
3	IDBSCAN	Boundary objects	$O(n\ log_m^n)$
4	LDBSCAN	Leader based clustering	$O((n + k)^2)$
5	FDBSCAN	Points representation technique	$O(n \log n)$
6	RDBSCAN	Set Theory	$O((n + k)^2)$
7	TIDBSCAN	Triangle Inequality	$O(n^3)$

Table 4: Density based algorithms and time complexity

S. No.	Algorithm	Preprocessed	Evaluation	Subspace extraction	cluster
1	DBSCAN	No	Samples within radius	All subspaces within threshold	Uniform cluster
2	DENCLUE	Yes	Analytical sum	Kernel functions	Arbitrary cluster shapes
3	GDBSCAN	Yes	Samples within radius	All subspaces within threshold	Uniform cluster
4	PDBSCAN	Yes	Samples within radius	All subspaces within threshold	Uniform cluster
5	SDBSCAN	Yes	Samples within radius	All subspaces within threshold	Clustering based on sampling theorem
6	SNNDBSCAN	No	Links of samples within radius	Uniform spaces	Arbitrary cluster shapes
7	IDBSCAN	Yes	Samples within radius	All subspaces within threshold	Arbitrary cluster shapes

8	STDBSCAN	Yes	Samples within radius with density	All subspaces within threshold	cluster shapes based on temporal data

Table 5: Density based algorithms and its features

Conclusion

Clustering is one of the Data Mining methods useful to group input data of different sizes into smaller clusters based on some parameters. Clustering of data can be done in many ways, density based clustering is one of the method to form clusters based on density of data samples and group high density data samples to clusters. In this paper, we conducted detailed survey on density based clustering techniques, applications, limitations, types of data used in clustering, delimitations, various sizes of data used, space occupied by each clustering technique, and running time of each technique.

References

Alireza, L., P., Daneshpour, N., "DBHC: A DBSCAN-based hierarchical clustering algorithm", Data & Knowledge Engineering, Vol. 135, September 2021, 101922.

Lulli, A., Dell Amico, M.; Michiardi, P., Ricci, L., "NG-DBSCAN: Scalable density based clustering for arbitrary data", In Proceedings of the VLDB Endow, New-Delhi, India, Vol. 10, pp. 157–168, 2016.

Yang, Y., Qian, C., Li, H., Gao, Y., Wu, J., Liu, C.-J., Zhao, S., "An efficient DBSCAN optimized by arithmetic optimization algorithm with opposition-based learning", The Journal of Supercomputing, vol. 78, pp.1–39, 2022.

Li, M., Bi, X., Wang, L., Han, X, "A method of two-stage clustering learning based on improved DBSCAN and density peak algorithm", Computer Communications, Vol. 167, pp. 75-8, 2021.

Atwa, W., Almazroi, A.A., "Active Selection Constraints for Semi-supervised Clustering Algorithms", International Journal of Information Technology and Computer Science, vol. 6, pp. 23-30, 2020.

Qasim, R., Bangyal, W.H., Alqarni, M.A., Almazroi, A.A., "A Fine-Tuned BERT Based Transfer Learning Approach for Text Classification", Journal of Healthcare Engineering, Vol. 2022, pp. 1-17, 2022.

Liu, X., Yang, Q., He, L., "A novel DBSCAN with entropy and probability for mixed data", Cluster Computing, pp. 1313–1323, vol. 20, 2017.

Kim, J.-H., Choi, J.-H., Yoo, K.-H., Nasridinov, A., "AA-DBSCAN: An approximate adaptive DBSCAN for finding clusters with varying densities", The Journal of Supercomputing, vol. 75, pp. 142–169, 2018.

Ablel-Rheem, D.M., Ibrahim, A.O., Kasim, S., Almazroi, A.A., Ismail, M.A., "Hybrid feature selection and ensemble learning method for spam email classification", International Journal Advanced Trends in Computer Science and

Engineering, vol. 9, no. 1.4, pp. 217–223, 2020.

Giri, M., Jyothi, S., "Big Data Collection and Correlation Analysis of Wireless Sensor Networks Yielding to Target Detection and Classification", In: Chaki, N., Cortesi, A., Devarakonda, N. (eds) Proceedings of International Conference on Computational Intelligence and Data Engineering. Lecture Notes on Data Engineering and Communications Technologies, vol 9. Springer, Singapore, 2018.

Masud, A., Huang, J.Z., Zhong, M., Fu, X., "Generate pair wise constraints from unlabeled data for semi-supervised clustering", Data & Knowledge Engineering, Vol. 123, 101715, 2019.

Qian, L., Plant, C., Bohm, C., "Density-Based Clustering for Adaptive Density Variation", 2021 IEEE International Conference on Data Mining (ICDM), 2021, pp. 1282-1287.

Campello, R.J., Moulavi, D., Zimek, A., Sander, J., "Hierarchical Density Estimates for Data Clustering Visualization and Outlier Detection", ACM Transaction on Knowledge Discovery from Data, vol. 10, no. 1, pp. 1-51, 2015.

Adeniyi, D.A., Wei, Z., Yongquan, Y., "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applications on Computer Informatics, vol. 12, no. 1, pp. 90-108, 2016.

Gallego, C.E.V., Comendador, V.F.G., Nieto, F.G.S., Martinez, M.G., "Discussion On Density-Based Clustering Methods Applied for Automated Identification of Airspace Flows", 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), pp. 1-10, 2018.

Campello, R. J. G. B., Moulavi, D., Zimek, A., Sander, J., "Hierarchical Density Estimates for Data Clustering Visualization and Outlier Detection", ACM Transaction on Knowledge Discovery from Data, vol. 10, no. 1, pp. 1-51, 2015.

Hou J, Gao H, Li X, "DSets-DBSCAN: a parameter-free clustering algorithm", IEEE Trans Image Process, vol. 25, no. 7, pp. 3182–3193, 2016.

Kim J, Lee W, Song JJ, Lee SB, "Optimized combinatorial clustering for stochastic processes", Cluster Computing, vol. 20, no. 2, pp. 1135–1148, 2017.

Smieja, M., Geiger, B.C., "Semi-supervised cross-entropy clustering with information bottleneck constraint", Information Science, vol. 421, pp. 254–271, 2017.

Masud, M.A., Huang, J.Z., Wei, C., Wang, J., Khan, I., Zhong, M., "I-nice: A new approach for identifying the number of clusters and initial cluster centers", Information Science, vol. 466, pp. 129–151, 2018.

Starczewski, A., Cader, A., "Grid-based approach to determining parameters of the DBSCAN algorithm", IC: Artificial Intelligence and Soft Computing (ICAISC), Springer, pp. 555–565, 2020.

Zhang, Y., Wang, X., Li, B., Chen, W., Wang, T., Lei, K., "DBOOST: a fast algorithm for DBSCAN-based clustering on high dimensional data", Pacific Asia Conference on Knowledge Discovery and Data

Mining, Springer, pp. 245–256, 2016.

Lv, Y., Ma, T., Tang, M., Cao, J., Tian, Y., Dheloan, A.A., "An efficient and scalable density-based clustering algorithm for datasets with complex structures", Neuro computing, vol. 171, No. C, pp. 9-22, 2016.

Kumar, K.M., Reddy, A.R.M., "A fast DBSCAN clustering algorithm by accelerating neighbor searching using groups method", Pattern Recognition, vol. 58, pp. 39–48, 2016.

Li, M., Bi, X., Wang, L., Han, X., "A method of two-stage clustering learning based on improved DBSCAN and density peak algorithm", Computer Communications, vol. 167 pp. 75–84, 2021.

Starczewski, A., Cader, A., "Determining the EPS parameter of the DBSCAN algorithm", IC: 18th International Conference on Artificial Intelligence and Soft Computing (ICAISC), Springer, pp. 420–430, 2019.