Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

Extensive Analysis of Clustering Algorithm for Large Datasets Using Density-Based Clustering and Swarm Intelligence

Rajendra DevidasMandaokar

SATI, Vidisha, MP

rdmandaokar@rediffmail.com

Dr. Shailesh Jaloree

SATI, Vidisha, MP

Shailesh_jaloree@rediffmail.com

Abstract

Clustering is the approach of unsupervised learning. The clustering process efficiently handles large scale dataset and maintains the purity of clusters—the density-based clustering algorithm can manage large scale dataset. The approach of clustering faces problems correctness and computational overhead.This paper improves the DBSCAN algorithm with partial probability function, increases cluster correctness, and reduces computational overhead. The applied swarm intelligence algorithm on a density-based algorithm manages the different clusters of clusters to merge with the same centres and EPS point. The process of particle swarm intelligence also reduces the noise and boundary value of data points and increases core points' value. This paper also studies stream data clustering. The stream data clustering modified the MCM algorithm with threshold function, the modified MCM Handel the evolved feature concept. These all algorithms were implemented in MATLAB software and applied high-dimension datasets such as a flame spiral pathway. The results' analysis indicates that the modified algorithm improves the correctness of cluster data approx. 2-4%.

KEYWORDS: Clustering, Density, Distance, DBSCAN, FCM, MCM, Large dataset, MATLAB.

Introduction

The advancement of technology and generation of high-speed data has multivariant. Multivariant data processing is very challenging due to unstructured and unformatted[1, 2]. The demands of automation required pattern analysis of stored data for further action of the task. For the analysis of pattern applied various data mining and machine learning algorithm[3, 4, 5]. Data mining offers various algorithms such as clustering, classification, and hybrid algorithms to analyze patterns and large-scale grouping data[6, 7, 8]. Theclustering algorithm is better instead of the classification algorithm for a large dataset. The process of clustering handles unknown attribute of the dataset on behaviors of iteration[9, 10, 11]. The clustering algorithm's processing categorizes into different categories such partition-based, density-based, hierarchical based and micro clustering algorithm[12, 13, 14]. The partition and density-based clustering algorithm are most appropriate for large dataset. The partition-based clustering algorithm is the umbrella of various clustering algorithms such ask-means, k-mode, FCM and others algorithm[15, 16, 17, 18]. The K-means and FCM is a very famous algorithm for the large dataset. The major problem with partition

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

clustering is data divergence. The density-based clustering algorithm overcomes these limitations of the partition-based clustering algorithm[19]. The density-based clustering algorithm diverse data with minimum points decreased the noise of data and improved the clustering algorithm's performance. The density-based clustering algorithms have many variants such as DBSCAN, IDBSCAN and other algorithm modified with optimization algorithm[20, 21]. Density-based clustering algorithms such as DBSCAN and OPTICS regard clusters as dense regions separated by sparse areas[22]. They are widely used because of their inherent abilities to deal with arbitrarily shaped clusters, automatically discover the number of clusters and identify the noise. The clustering process of density-based algorithms can be split into two steps[23, 24, 25]. First, we define a procedure for estimating each sample's density and apply it to identify the samples within dense regions. Second, we search for the connected dense regions and assign them to the same cluster. The major advantage of density-based clustering over other clustering algorithm is the handling of noise capabilities. The other clustering algorithm cannot handle the noise data and deform the process of clustering[26, 27]. The density and distance play a major role in clustering algorithms. The distance function estimate similarity of cluster center through objects, and density measures the volume of data. the density and distance influence the correctness of the cluster validation process[28]. The correctness of the cluster depends on the purity of data to the formation of the cluster. Some part of the data treated as noise and outlier. The value of noise and outlier decline the performance of the clustering algorithm. The various authors and research scholar reported the minimization of noise and correct selection of centers applied swarm intelligence algorithms[29, 30]. The swarm intelligence algorithm improves the efficiency of the clustering algorithm[31]. The swarm intelligence algorithm processing based on two modes of operation, such as single objective function and multi-objective function. The multi-objective function changes the clustering algorithm's working process and known as the name of swarm algorithm such as SCA (swarm clustering algorithm). KANT (ant clustering algorithm). The SCA and KANT algorithms handle large-scale datasets such as spatial image data and real-time big data processing to analyze pattern. The majority of existing algorithms strive to determine a set of cluster centers, and it is challenging for them to deal with a cluster of arbitrary shape[32]. The clustering algorithm's other challenge is the number of iteration and formation of valid shape and pattern during the intermediate cluster validation. Most clustering algorithms handle data better, but in the case of stream data, the clustering process suffered. The stream data is a continuous transaction of any function. All the process of automation required the analysis of data. Nowthe stream data also needs the formation and analysis of cluster. The density-based clustering is suitable for stream data clustering. The major challenges in-stream data clustering is evolving new features[33, 34]. The evolved new features during the process of clustering treat as noise and outlier of data. Theevolved new features are core attribute of data and cannot participate in pattern and cluster analysis. This paper mainly focuses on large scale data clustering algorithm based on the function of density and distance. The density-based algorithms such as DBSCAN, IDBSCAN and others improved algorithm based on swarm intelligence. The process of analysis also focuses on partition-based clustering algorithm and stream data clustering[35]. The rest of the paper describes as in section 2, describe related work in the area of large-scale data clustering. In section 3. Describe the process of the methodology of clustering. In section 4. Describe the experimental analysis and finally discuss the conclusion and future work in section 5.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

2. Related Work

The diversity of clustering algorithms attracts the research scholar for continuous improvement of pattern extraction and data analysis. Various authors and research scholar contribute in the area of clustering algorithm with various optimization algorithms. Some contribution of authors are described here.

Thrun, Michael C. Et al. [1] For high-dimensional datasets in which bunches are framed by the two DDS, many grouping calculations neglect to recognize these groups effectively. This is shown for 32 grouping calculations utilizing a set-up of datasets which intentionally present complex DDS challenges for bunching. To improve the design finding and grouping in high-dimensional DDS datasets, PBC is presented. The concurrence of projection and bunching permits to investigate DDS through a geographical guide. This empowers to gauge, first, if any bunch inclination exists and, second, the assessment of the quantity of groups. A correlation showed that PBC is consistently ready to track down the right bunch structure, while the exhibition of the best of the 32 grouping calculations fluctuates relying upon the dataset. Luo, Wenjian Et al. [2] a novel proficient multitude grouping calculation named SCA2 is talked about, which expands SCA as far as three perspectives: the outspread premise work network is embraced as the substitute model to lessen the time intricacy; there are k pioneers for every molecule, and the molecule may follow one of them to diminish misdirecting; and an improved-on system is utilized to refresh the situation of every molecule. The exhibition of SCA2 on various sorts of engineered and genuine world datasets was contrasted and the presentation of four old style calculations, SCA just as a PSO-based grouping calculation. The test results exhibit that SCA2 is more serious. Bulut, Hasan Et al. [3] The microarray innovation empowers the examination of the quality articulation information and the comprehension of the significant natural cycles in an efficient way. Creators have built up an efficient bunching plan for microarray quality articulation information dependent on relationship-based component choice, insect-based grouping, fluffy c-implies calculation and a novel pile combining heuristic. The calculation uses the element choice calculation to beat the high-dimensionality issue experienced in bioinformatics area. In view of broad observational investigation on microarray information, bunching nature of the subterranean insect-based grouping calculation is improved with the utilization of fluffy c-implies calculation and stores consolidating heuristic. The exhibition of the examined grouping plan is contrasted and k-implies, PAM calculation, CLARA, self-sorting out map, progressive bunching, troublesome investigation grouping, self-putting together tree calculation, half and half various leveled grouping, agreement grouping, AntClass calculation and fluffy c-implies grouping calculations. The trial results demonstrate that the examined grouping plan yields better execution in bunching disease quality articulation information.

Jang, Jennifer Et al. [4] DBSCAN is an old-style thickness based bunching strategy with enormous functional importance. Be that as it may, DBSCAN verifiably needs to register the experimental thickness for each example point, lead-ing to a quadratic most pessimistic scenario time intricacy, which is too delayed on enormous datasets. Creators talked about DBSCAN++, a basic adjustment of DBSCAN which just requires figuring the densities for a picked subset of focuses. Creators show exactly that, contrasted with conventional DBSCAN, DBSCAN++ can give serious execution as well as added strength in the transfer speed hyperparameter while taking a small part of the runtime. Creators likewise present measurable consistency ensures showing the compromise between computational expense and assessment rates. Shockingly, in a measured way, creators can appreciate a similar assessment rates while bringing down computational expense, showing that DBSCAN++ is a sub-quadratic calculation that achieves minimax ideal rates for level-set assessment, a quality

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

that might be of autonomous premium. Thrun, Michael C. Et al. [5] The DBS is an adaptable and strong grouping system that comprises of three autonomous modules: swarm-based projection, high-dimensional information perception and portrayal guided bunching. The primary module is the boundary free projection strategy Pswarm, which abuses ideas of self-association and development, game hypothesis, and multitude knowledge. The subsequent module is a boundary free high-dimensional information representation strategy called geographical guide. It utilizes the summed-up U-grid, which empowers to appraise first, if any group propensity exists and second, the assessment of the quantity of bunches. The third module offers a grouping strategy which can be checked by the representation and the other way around. Benchmarking w.r.t. regular calculations exhibited that DBS can out-perform them. A few applications showed that group structures given by DBS are significant. Praiseworthy, a grouping of overall nation related information w.r.t. the COVID-19 pandemic is introduced here. Code and information is made accessible by means of open source. Amini, Amineh Et al. [6] creators talked about a thickness-based grouping calculation for IoT streams. The technique has quick handling time to be appropriate continuously use of IoT gadgets. Test results show that the examined approach acquires excellent outcomes with low calculation time on genuine and manufactured datasets.

Cai, Zihao Et al. [7] Clustering is an old-style research field because of its expansive applications in information mining like feeling identification, occasion extraction and subject disclosure. It plans to find inherent examples which can be shaped as groups from an assortment of information. Significant progress have been made by the DBSCAN and its variations. Nonetheless, there is a significant limit that current thickness-based calculations experience the ill effects of direct association issue, where they perform inadequately to segregate target bunches which are "associated" by a couple of information focuses. In addition, the boundary setting and the time cost make it difficult to be all around adjusted in monstrous information investigation. To address these issues, creators examined a novel versatile thickness based spatial bunching calculation called Ada-DBSCAN, which comprises of an information block splitter and an information block consolidation, facilitated by nearby grouping and worldwide bunching. Creators direct broad tests on both artificial and genuine world datasets to assess the viability of Ada-DBSCAN. Test results show that our calculation obviously beats a few in number baselines in both grouping exactness and human assessment. Furthermore, Ada-DBSCAN shows significant improvement of efficiency contrasted and DBSCAN. Gaonkar, Manisha Naik Et al. [8] Emergence of current methods for logical information assortment has brought about enormous scope aggregation of information relating to different fields. Ordinary data set questioning techniques are insufficient to extricate valuable data from enormous information banks. Group examination is an essential technique for information base mining. It is either utilized as an independent instrument to get knowledge into the appropriation of an informational index or as a pre-preparing venture for different calculations working on the distinguished bunches. Practically the entirety of the notable grouping calculations requires input boundaries which are difficult to decide however affect the bunching result. Besides, for some, genuine informational collections there doesn't exist a worldwide boundary setting for which the consequence of the grouping calculation portrays the characteristic bunching structure precisely. DBSCAN is a base calculation for thickness-based grouping methods. This paper gives an overview of thickness-based grouping calculations with the examined improved calculation that naturally chooses the information boundaries alongside its execution and correlation with the current DBSCAN calculation. The exploratory outcomes shows that the examined calculation can recognize the groups of changed thickness with various shapes and sizes from huge measure of information

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

which contains clamor and exceptions, requires just one information boundaries and gives better yield then the DBSCAN calculation.

Kokate, Umesh Et al. [9] different information stream strategies and calculations are inspected and assessed on standard engineered information streams and genuine information streams. Thickness miniature bunching and thickness network-based grouping calculations are talked about and near examination as far as different inside and outer grouping assessment strategies is performed. It was seen that a solitary calculation can't fulfill all the exhibition measures. The exhibition of these information stream bunching calculations is area explicit and requires numerous boundaries for thickness and clamor limits. Dharni, Chetan Et al. [10] Clustering is a significant instrument which has seen a dangerous development in Machine Learning Algorithms. DBSCAN bunching calculation is quite possibly the most essential strategies for grouping in information mining. DBSCAN has capacity to discover the groups of variable sizes and shapes and it will likewise distinguish the clamor. The two significant boundaries Eps and MinPts are needed to be inputted physically in DBSCAN calculation and on the premise these boundaries the calculation is determined like number of bunch, un-grouped occasions just as erroneously bunched cases and furthermore assess the exhibition on the essential of boundaries choice and ascertain the time taken by the datasets. Exploratory assessment based on various datasets in ARFF design with assistance of WEKA device which shows that nature of bunches of our talked about calculation is effective in grouping result and more exact. This improved work on DBSCAN have utilized in a huge extension.Nguyen, Hai-Long Et al. [11] with the development of innovation, numerous applications create colossal measures of information streams at extremely fast. Models incorporate organization traffic, web click transfers, video reconnaissance, and sensor organizations. Information stream mining has become a hot examination point. Its will likely concentrate covered up information/designs from ceaseless information streams. Not at all like conventional information mining where the dataset is static and can be more than once read ordinarily, information stream mining calculations face numerous difficulties and need to fulfil requirements like limited memory, single-pass, constant reaction, and idea float identification. This paper presents a far-reaching study of the cutting-edge information stream mining calculations. It identifies mining limitations and talked about an overall model for information stream mining, and portrays the connection between conventional information mining and information stream mining.

Heidari, Safanaz Et al. [12] creators have endeavoured to present another calculation for grouping enormous information with shifted thickness utilizing a Hadoop stage running MapReduce. The fundamental thought of this exploration is the utilization of neighbourhood thickness to discover each point's thickness. This technique can dodge the circumstance of interfacing groups with changing densities. The examined calculation is carried out and contrasted and different calculations utilizing the MapReduce worldview and shows the best fluctuating thickness grouping capacity and adaptability. Sleeman, William C. Et al. [13] creators talked about the principal compound system for managing multi-class large information issues, tending to simultaneously the presence of different classes and high volumes of information. Creators examined to investigate the example level challenges in each class, prompting understanding what causes realizing troubles. Creators install this data in well-known resampling calculations which takes into account educational adjusting of numerous classes. Creators examined a productive execution of the talked about calculation on Apache Spark, including a novel form of SMOTE that beats spatial restrictions in appropriated conditions of its archetype. Li, Hao Et al. [14] creators present a novel methodology for recognizing neighbourhood high-thickness tests using the innate properties of the NNG. In the wake of utilizing the thickness assessor to channel commotion tests, the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

talked about calculation ADBSCAN in which represents Adaptive plays out a DBSCAN-like bunching measure. The test results on fake and genuine world datasets have shown the critical execution improvement over existing thickness-based grouping calculations.

3. Methodology of Clustering Algorithm

This section describes the methodology of cluster formation. The formation of cluster based on density and distance function. The density belongs to DBSCAN clustering algorithm and distance belongs to partition-based clustering algorithms. The density-based clustering algorithms have low computational value and very efficiently handle the large data scale. The processing of clustering depends on two points radius of cluster(EPS) and nearest neighbors' points (Min-points). Here describe the DBSCAN clustering algorithm and improved DBSCAN clustering algorithm, further describe the multi-class miner algorithm and modified MCM algorithm is called PMCM[1, 2].

DBSCAN ALGORITHM

The DBSCAN algorithms depends on three factors such as core points, border points and noise points. The core points are interior of cluster. The process of algorithm handles with two parameters Eps and MinPts. Border points is also core points of cluster but lies on boundary of clusters. The noise points are any data points that is neither a core point nor a border point[3, 4, 5].

Steps of algorithm

1. Labelling of core, border or noise points
2. Discard noise points
3. Estimate all core points within Eps.
4. Separate the clusters
5. Assign border point of each cluster

PPDBSCAN algorithm

The modified DBSCAN algorithm applied the function of partial probability. The partial probability merges the two adjacent cluster based on probability function. The probability function reduces the boundary and noise value of data points as:

P1 and P2 are distributed in different regions RP1 and RP2 respectively:

Eps(P1) is the Eps of P1 and Eps(P2) means Eps of P2 and partial probability of P1 and P2 is

PPMEps(P1, P2).

Now distance of two objects Pi and Qi

$$dist = \frac{\sum_{i=1}^{p1}\sum_{j=1}^{p2} distance\ (p1,Qj)}{NP1\ XNp2} \leq \text{PPMEps(P1,P2)}$$

The processing of PPDBSCAN algorithm

Input: dataset D, Minpts, Eps
Output: set of clusters
Estimate the probability of data points with entropy
For each partial probability PP in the dataset D do
Put points in adjacent probability

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

End for
C1=PP(cluster)
Transform the value of probability to shift by size/2
C2=PPDBSCAN(PP)
C1=|c1| total number of distinct data points in C1
C2=|C2| total number of distinct data points in C2
M=zeros(C1+1) X(C2+1)
M is core points of Clusters
For each point P in D do
C1-cluester=cluester(P1) in c1
C2-cluester=cluester(P2) in C2
M[c1-cluester] [c2-cluester] =1
End for

$$if\ M_{ij} \geq Minpts\ and\ M_{i,j} \geq Minpts\ then$$

Merge C1 and C2
End if
End for
Return cluster

MCM ALGORITHM

The handling of stream data clustering is very difficult to normal clustering algorithm and density-based clustering algorithm. The density-based clustering enhances with threshold function and applied on stream data for the process of clustering. The threshold function describes as

$$Th_{pt} = \log_{\lambda} \frac{\alpha}{\alpha} - N(1 - 2^{-\lambda})$$

Here Th is the value of threshold for the selection of new feature points of stream data. the MCM is main algorithm of stream data clustering. The MCM algorithm incorporate with density-based clustering and formed cluster for stream data.

Process of algorithm

Input: stream data, Mints, $.\lambda$ and α

Output: clusters

1.Define the threshold as derivation

2. T_p=0

3. Process stream data do

4. Read data point x form data stream

5. Estimate nearest mini-cluster to x

6. If dist(x, centers) <rmcm then

7. Merge x to MCM

8. Else

9. Map the new data points x to the MCM

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

11 Update MCM

12. End if

13. Return cluster.

PARTICLE SWARM OPTIMIZATION (PSO)

Particle swarm optimization resolve the problem of partition K-means clustering algorithm. The problem of K-means algorithm is minimization of objective function of similarity index. The fitness function of particle swarm optimization converted into objective function of clustering algorithm. The particle swarm optimization algorithm is dynamic population based meta-heuristic function. The working of algorithm deals in two manners local and global ways. The local mapping deals with selection of data points and global maintain the objective function of clustering algorithm. The processing of PSO algorithm based on the concept of bird fork and movements of fish in wheel[1]. The velocity and direction of particle moving with problem space of data. the mapping of process of data space in velocity and direction shown in figure[3].

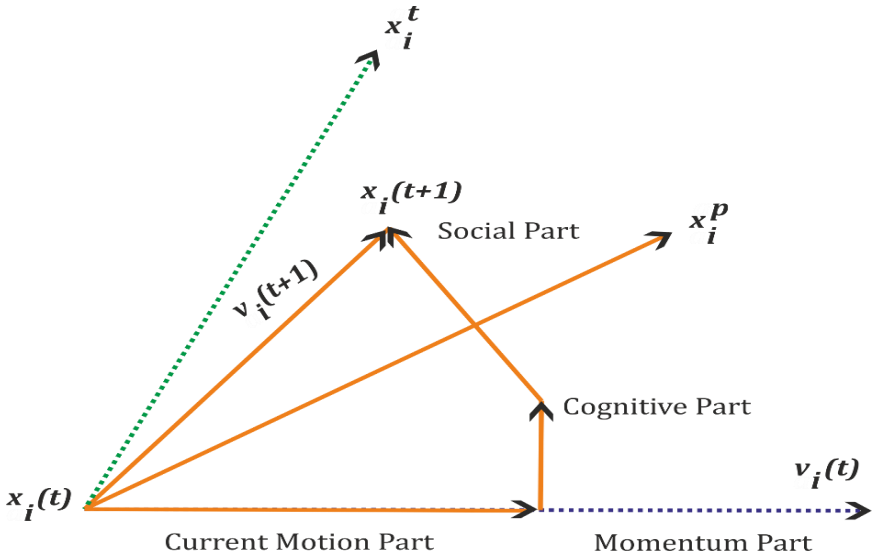

Figure 1: Represents the velocity and direction of particle in search space for the objective function.

the processing of particle swarm optimization algorithm is described here.

1. Define the population of particle as data points A
   (a) For i= 0 to M where M is maximum of particle
   (b) Initialize A[i]
2. Define the speed of particle
   (a) For i= 0 to M
   (b) Velocity[i]=0
3. Estimate particle in M
4. Reallocate the position that represents the data point of data samples
5. Generate search space D
6. Define memory o each particle
7. Compute the speed of particle
   Velocity[i]= W*Velocity[i]+R1*(Pbset[i]-M[i])+R2*(data points[h]-A[i])

   Where the range value of R is[0,1]

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

8. Estimate new position of particle
   A[i]=A[i]+velocity[i]
9. Measure current position of particle
   Pbest[i]=A[i]
10. Increment of counter
11. End

4. Experimental Analysis

To evaluate the performance of modified DBSCAN algorithm and modified MCM algorithm with density-based clustering implement in MATLAB software with version R2014(a). Thesystem configuration of device is windows 10 operating system, processor I7, RAM 16GB and HDD 1TB. For the analysis of methods applied three standard parameters such as correctness of cluster, elapsed time and error.

The correctness of cluster validates the purity of clustering process. The formula of correctness of cluster as

$$corectness = \frac{\sum_{i=1}^{k}(\left|c_i^d\right| /|ci|)}{k} X100 \dots\dots\dots\dots\dots\dots\dots (1)$$

The error of cluster indicates the factor of standard deviation as

$$CE = \sum_{i} \sum_{x \epsilon Ci} (x - mi)^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

Time complexity is major factor in analysis of cluster validation. The formulation of cluster validation as

$$T(n) = T(n-1) + 1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

DATASET

The performance of clustering algorithms measure on six data set. The applied all dataset is high dimension and large size. The name of dataset is flame dataset, spiral dataset, path-based dataset and another dimension dataset name as S1, S2, S3 and S4. The source of dataset is UCI machine learning repository and KEEL.

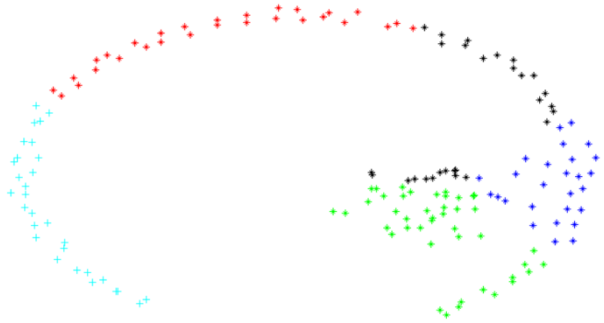

Figure 2: window show that the output of spatial data mining cluster and here loads the dataset on the basis of input field of EPS value is 0.2, here hit the DBscan technique than in the GUI interface.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

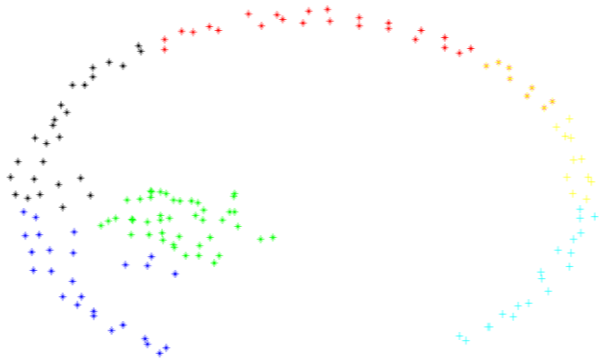

Figure 3: window show that output of the spatial data mining cluster and here loads the dataset on the basis of input field of EPS value is 0.2, here hit the IDBscan technique than in the GUI interface.

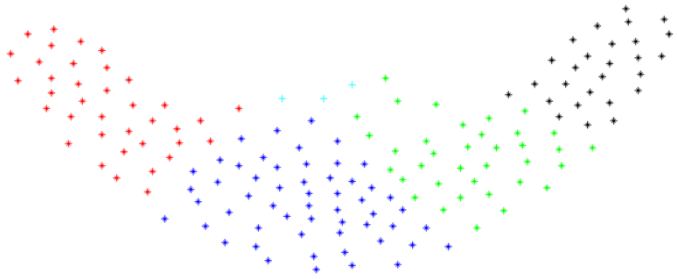

Figure 4: window show that output of the spatial data mining technique and here loads the flame dataset on the basis of input field of EPS value is 0.5, here hit the K-Means technique than in the GUI interface.

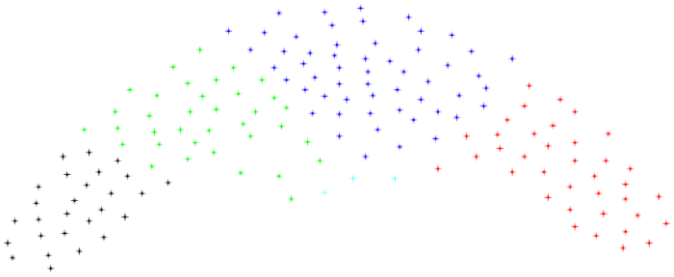

Figure 5: window show that parameters elapsed time, correctness and error value numeric result and here loads the flame dataset on the basis of input field of EPS value is 0.5, here hit the MCM technique than in the GUI interface.

Table 1: Comparative Analysis of DBscan, IDBscan, K-Means, FCM, MCM and PMCM technique using value of EPS is 0.2 and Path dataset with given parameters Elapsed Time, Correctness, Error Value.

	DBscan[1]	IDBscan[4]	K-Means[18]	FCM[3]	MCM[16]	PMCM	Dataset
Elapsed Time	13.63	12.57	10.46	10.03	10.58	9.95	Path Dataset
Correctness	97.19	97.05	93.24	97.15	96.35	98.01	

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

Error Value	2.50	2.90	1.90	1.67	1.54	1.49	

Table 2: Comparative Analysis of DBscan, IDBscan, K-Means, FCM, MCM and PMCM technique using value of EPS is 0.5 and Dim dataset with given parameters Elapsed Time, Correctness, Error Value.

	DBscan[1]	IDBscan[4]	K-Means[18]	FCM[3]	MCM[16]	PMCM	Dataset
Elapsed Time	18.65	15.37	15.44	12.75	11.33	11.12	Dim Dataset
Correctness	88.84	88.35	86.51	89.92	90.25	91.21	
Error Value	3.45	3.69	2.87	2.62	2.87	2.88	

Table 3: Comparative Analysis of DBscan, IDBscan, K-Means, FCM, MCM and PMCM technique using value of EPS is 0.7 and Flame dataset with given parameters Elapsed Time, Correctness, Error Value.

	DBscan[1]	IDBscan[4]	K-Means[18]	FCM[3]	MCM[16]	PMCM	Dataset
Elapsed Time	14.85	15.15	16.67	13.41	14.35	12.51	Flame Dataset
Correctness	90.64	91.24	90.07	91.09	89.27	88.69	
Error Value	4.41	5.65	3.67	3.11	3.46	3.07	

Table 4: Comparative Analysis of DBscan, IDBscan, K-Means, FCM, MCM and PMCM technique using value of EPS is 0.9 and Spiral dataset with given parameters Elapsed Time, Correctness, Error Value.

	DBscan[1]	IDBscan[4]	K-Means[18]	FCM[3]	MCM[16]	PMCM	Dataset
Elapsed Time	17.67	16.45	16.36	13.86	14.86	13.51	Spiral Dataset
Correctness	78.35	79.78	78.35	82.55	82.74	83.56	
Error Value	6.68	5.42	4.98	4.25	4.14	3.98	

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

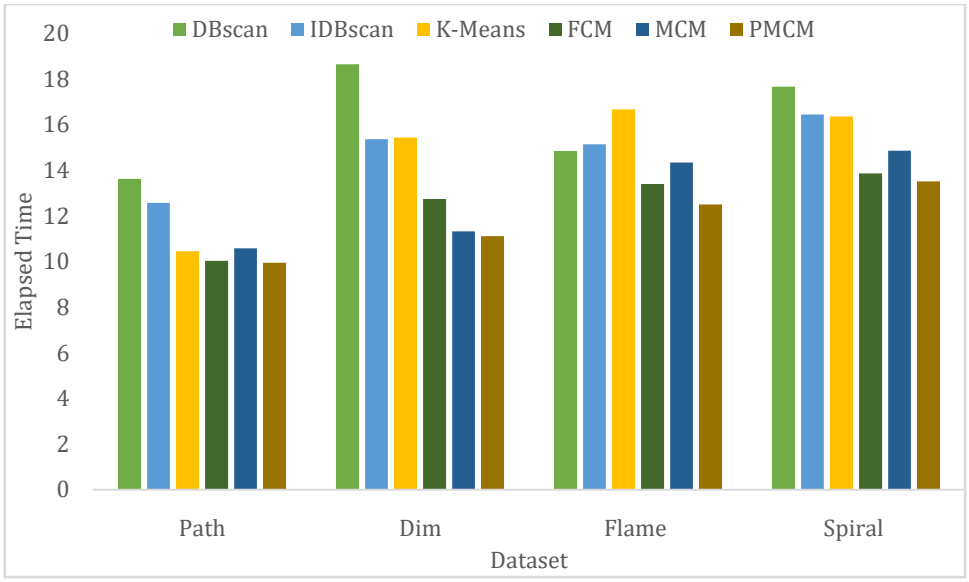

Figure 6: Comparative analysis of Elapsed Time using DBscan, IDBscan, K-Means, FCM, MCM and PMCM technique using value of EPS is 0.2, 0.5, 0.7, 0.9 and Path, Dim, Flame, Spiral dataset.

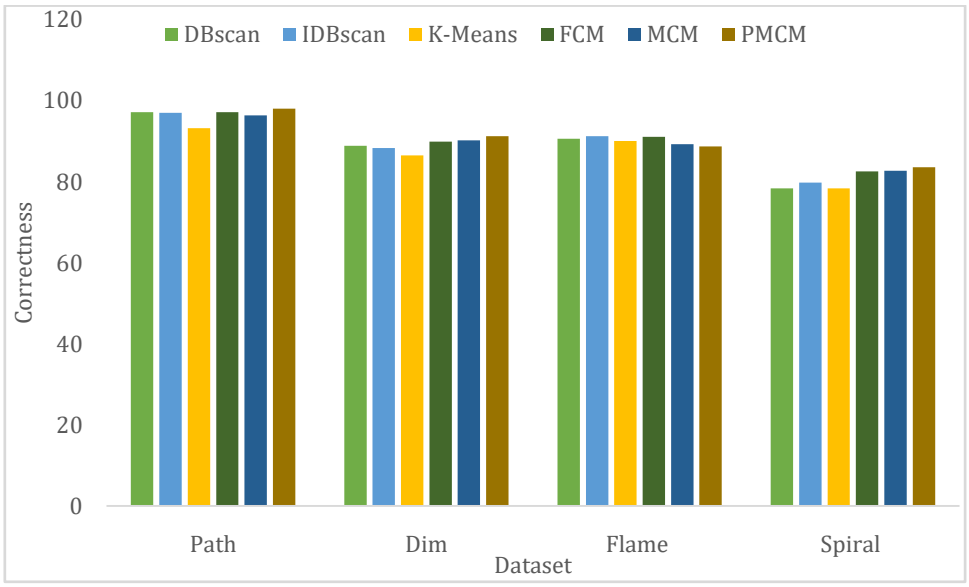

Figure 7: Comparative analysis of correctness using DBscan, IDBscan, K-Means, FCM, MCM and PMCM technique using value of EPS is 0.2, 0.5, 0.7, 0.9 and Path, Dim, Flame, Spiral dataset.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

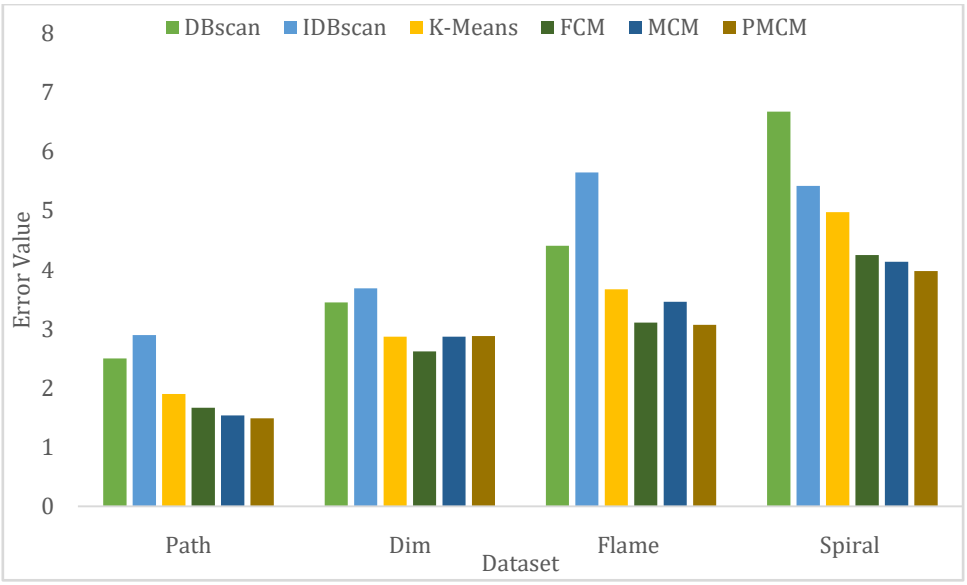

Figure 8: Comparative analysis of error value using DBscan, IDBscan, K-Means, FCM, MCM and PMCM technique using value of EPS is 0.2, 0.5, 0.7, 0.9 and Path, Dim, Flame, Spiral dataset.

5. Conclusion & Future Work

In this paper, we propose partial probability based improved DBSCAN algorithm for large dataset. The proposed algorithm is very efficient in terms of cluster correctness and computational time. The partial probability measures the correct Eps value of adjacent region for the nearest data points and minimize the noise and boundary data. the process of study also focusses on stream data clustering. The stream data clustering modified with the threshold function. The threshold function reduces the gap of cluster distance and form correct cluster. The threshold-based function also handles the problem of new feature evaluation. The evolved new features mapped with nearest cluster and reduces the rate of error. Both proposed algorithms compare with existing algorithm such as DBSCAN, K-means, FCM and MCM. The Proposed algorithm can effectively analyse the cluster for large dataset. The Improved algorithm is scalable than the density-based algorithm as it works on splitting dataset instead of working on whole dataset. The dataset having the total sum of both instances and attribute are more, the number of formed clusters as well as incorrectly clustered instances is also more. For future work, we plan to extend our algorithm to support distributed computing to fully explore the idea of data splitting and merging based on uniform data distribution, making it better-adapted to massive distributional data analysis. Another research direction is to exploit the ensemble clustering strategy to further improve our PP-DBSCAN.

References

[1]. Thrun, Michael C., and Alfred Ultsch. "Using Projection-Based Clustering to Find Distance-and Density-Based Clusters in High-Dimensional Data." *Journal of Classification* (2020): 1-33.

[2]. Luo, Wenjian, Wenjie Zhu, Li Ni, Yingying Qiao, and Yigui Yuan. "SCA2: Novel Efficient Swarm Clustering Algorithm." *IEEE Transactions on Emerging Topics in Computational Intelligence* (2020).

[3]. Bulut, Hasan, Aytuğ Onan, and Serdar Korukoğlu. "An improved ant-based algorithm based on heaps merging and fuzzy c-means for clustering cancer gene expression data." *Sādhanā* 45, no. 1 (2020): 1-17.
[4]. Jang, Jennifer, and Heinrich Jiang. "DBSCAN++: Towards fast and scalable density clustering." In *International Conference on Machine Learning*, pp. 3019-3029. PMLR, 2019.
[5]. Thrun, Michael C., and Alfred Ultsch. "Swarm intelligence for self-organized clustering." *Artificial Intelligence* 290 (2021): 103237.
[6]. Amini, Amineh, HadiSaboohi, Teh Ying Wah, and TututHerawan. "A fast density-based clustering algorithm for real-time internet of things stream." *The Scientific World Journal* 2014 (2014).
[7]. Cai, Zihao, Jian Wang, and Kejing He. "Adaptive density-based spatial clustering for massive data analysis." *IEEE Access* 8 (2020): 23346-23358.
[8]. Gaonkar, Manisha Naik, and Kedar Sawant. "AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset." *International Journal on Advanced Computer Theory and Engineering* 2, no. 2 (2013): 11-16.
[9]. Kokate, Umesh, Arvind Deshpande, Parikshit Mahalle, and Pramod Patil. "Data stream clustering techniques, applications, and models: comparative analysis and discussion." *Big Data and Cognitive Computing* 2, no. 4 (2018): 32.
[10]. Dharni, Chetan, and Meenakshi Bnasal. "An improvement of DBSCAN Algorithm to analyze cluster for large datasets." In *2013 IEEE international conference in MOOC, innovation and technology in education (MITE)*, pp. 42-46. IEEE, 2013.
[11]. Nguyen, Hai-Long, Yew-KwongWoon, and Wee-Keong Ng. "A survey on data stream clustering and classification." *Knowledge and information systems* 45, no. 3 (2015): 535-569.
[12]. Heidari, Safanaz, Mahmood Alborzi, Reza Radfar, Mohammad Ali Afsharkazemi, and Ali RajabzadehGhatari. "Big data clustering with varied density based on MapReduce." *Journal of Big Data* 6, no. 1 (2019): 1-16.
[13]. Sleeman IV, William C., and Bartosz Krawczyk. "Multi-class imbalanced big data classification on Spark." *Knowledge-Based Systems* 212 (2021): 106598.
[14]. Li, Hao, Xiaojie Liu, Tao Li, and Rundong Gan. "A novel density-based clustering algorithm using nearest neighbor graph." *Pattern Recognition* 102 (2020): 107206.
[15]. Azad, Ariful, Georgios A. Pavlopoulos, Christos A. Ouzounis, Nikos C. Kyrpides, and Aydin Buluç. "HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks." *Nucleic acids research* 46, no. 6 (2018): e33-e33.
[16]. Du, Mingjing, Shifei Ding, and Yu Xue. "A robust density peaks clustering algorithm using fuzzy neighborhood." *International Journal of Machine Learning and Cybernetics* 9, no. 7 (2018): 1131-1140.
[17]. Rodriguez, Mayra Z., Cesar H. Comin, Dalcimar Casanova, Odemir M. Bruno, Diego R. Amancio, Luciano da F. Costa, and Francisco A. Rodrigues. "Clustering algorithms: A comparative approach." *PloS one* 14, no. 1 (2019): e0210236.
[18]. Shakeel, P. Mohamed, S. Baskar, VR Sarma Dhulipala, and Mustafa Musa Jaber. "Cloud based framework for diagnosis of diabetes mellitus using K-means clustering." *Health information science and systems* 6, no. 1 (2018): 16.
[19]. Sinha, Debajyoti, Akhilesh Kumar, Himanshu Kumar, Sanghamitra Bandyopadhyay, and Debarka Sengupta. "dropClust: efficient clustering of ultra-large scRNA-seq data." *Nucleic acids research* 46, no. 6 (2018): e36-e36.
[20]. Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. "SCANPY: large-scale single-cell gene expression data analysis." *Genome biology* 19, no. 1 (2018): 15.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6368 - 6382
Received 25 April 2021; Accepted 08 May 2021.

[21]. Yang, Yan, and Hao Wang. "Multi-view clustering: A survey." *Big Data Mining and Analytics* 1, no. 2 (2018): 83-107.
[22]. Liu, Qidong, Ruisheng Zhang, Rongjing Hu, Guangjing Wang, Zhenghai Wang, and Zhili Zhao. "An improved path-based clustering algorithm." *Knowledge-Based Systems* 163 (2019): 69-81.
[23]. Patibandla, RSM Lakshmi, and N. Veeranjaneyulu. "Performance Analysis of Partition and Evolutionary Clustering Methods on Various Cluster Validation Criteria." *Arabian Journal for Science and Engineering* 43, no. 8 (2018): 4379-4390.
[24]. Riaz, Sumbal, Mehvish Fatima, Muhammad Kamran, and M. Wasif Nisar. "Opinion mining on large scale data using sentiment analysis and k-means clustering." *Cluster Computing* 22, no. 3 (2019): 7149-7164.
[25]. Zou, Quan, Gang Lin, Xingpeng Jiang, Xiangrong Liu, and Xiangxiang Zeng. "Sequence clustering in bioinformatics: an empirical study." *Briefings in bioinformatics* 21, no. 1 (2020): 1-10.
[26]. Deng, Tingquan, Dongsheng Ye, Rong Ma, Hamido Fujita, and LvnanXiong. "Low-rank local tangent space embedding for subspace clustering." *Information Sciences* 508 (2020): 1-21.
[27]. Huang, Shudong, Zhao Kang, and Zenglin Xu. "Self-weighted multi-view clustering with soft capped norm." *Knowledge-Based Systems* 158 (2018): 1-8.
[28]. Vandenbon, Alexis, and Diego Diez. "A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data." *Nature communications* 11, no. 1 (2020): 1-10.
[29]. Shaham, Uri, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. "Spectralnet: Spectral clustering using deep neural networks." *arXiv preprint arXiv:1801.01587* (2018).
[30]. Bloomfield, Nathaniel J., Nunzio Knerr, and Francisco Encinas-Viso. "A comparison of network and clustering methods to detect biogeographical regions." *Ecography* 41, no. 1 (2018): 1-10.
[31]. Wang, Jingwen, and Jie Wang. "Hierarchical Topic Clustering over Large Collections of Documents." In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, pp. 10-16. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2019.
[32]. Zhang, Zheng, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. "Binary multi-view clustering." *IEEE transactions on pattern analysis and machine intelligence* 41, no. 7 (2018): 1774-1782.
[33]. Gerlach, Martin, Beatrice Farb, William Revelle, and Luís A. Nunes Amaral. "A robust data-driven approach identifies four personality types across four large data sets." *Nature human behaviour* 2, no. 10 (2018): 735-742.
[34]. Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep clustering for unsupervised learning of visual features." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132-149. 2018.
[35]. Gallego, Antonio-Javier, Jorge Calvo-Zaragoza, Jose J. Valero-Mas, and Juan R. Rico-Juan. "Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation." *Pattern Recognition* 74 (2018): 531-543.