I. J. Computer Network and Information Security, 2021, 6, 70-82
Published Online December 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2021.06.06

An Optimized K-means with Density and Distance-Based Clustering Algorithm for Multidimensional Spatial Databases

K Laskhmaiah[1], S Murali Krishna[2], B Eswara Reddy[3]
[1] Department of Computer Science and Engineering, JNTUH, Hyderabad, Telangana, India
E-mail: klakshmaiah78@gmail.com

[2] Department of Information Technology, SVCE, Tirupati, Andhra Pradesh, India
E-mail: muralikrishna.s@svcolleges.edu.in

[3] Department of Computer Science and Engineering, JNTUA, Ananthapuram,Andhra Pradesh, India
E-mail: eswarcsejntua@gmail.com

Received: 25 March 2021; Revised: 12 May 2021; Accepted: 28 June 2021; Published: 08 December 2021

Abstract: From massive and complex spatial database, the useful information and knowledge are extracted using spatial data mining. To analyze the complexity, efficient clustering algorithm for spatial database has been used in this area of research. The geographic areas containing spatial points are discovered using clustering methods in many applications. With spatial attributes, the spatial clustering problem have been designed using many approaches, but non-overlapping constraints are not considered. Most existing data mining algorithms suffer in high dimensions. With non-overlapping named as Non Overlapping Constraint based Optimized K-Means with Density and Distance-based Clustering (NOC-OKMDDC), a multidimensional optimization clustering is designed to solve this problem by the proposed system and the clusters with diverse shapes and densities in spatial databases are fast found. Proposed method consists of three main phases. Using weighted convolutional Neural Networks (Weighted CNN), attributes are reduced from the multidimensional dataset in this first phase. A partition-based algorithm (K-means) used by Optimized K-Means with Density and Distance-based Clustering (OKMDD) and several relatively small spherical or ball-shaped sub clusters are made by Clustering the dataset in this second phase. The optimal sub cluster count is performed with the help of Adaptive Adjustment Factorbased Glowworm Swarm Optimization algorithm (AAFGSO). Then the proposed system designed an Enhanced Penalized Spatial Distance (EPSD) Measure to satisfy the non-overlapping condition. According to the spatial attribute values, the spatial distance between two points is well adjusted to achieving the EPSD. In third phase, to merge sub clusters the proposed system utilizes the Density based clustering with relative distance scheme. In terms of adjusted rand index, rand index, mirkins index and huberts index, better performance is achieved by proposed system when compared to the existing system which is shown by experimental result.

Index Terms: Spatial Data Mining, Overlapping, Weighted Convolutional Neural Networks (Weighted CNN) and clustering.

1. Introduction

Data clustering is very important field in data mining and knowledge discovery to unhide latent information in data, the popularity of density clustering methods is due to its ability to handle data with varied shapes and sizes clusters, does not require number of clusters in advance and handle well noise and outliers. Over the past 30 years, significantly, the mapping and analyzing of crime have evolved. With colored pins, city and precinct are utilized by many agencies to visualizing individual crime events and crime plagued areas in the beginning. Now- a- days computer-based techniques for exploring, visualizing, and explaining the occurrences of criminal activity, with the rapid advancement of technology have been essential. Exploration of the spatial distribution of crime has been facilitated which is one of the most influential tools and has been Geographical Information System (GIS) [1]. With other data that makes GIS so valuable is combined by it. Spatial Data Mining is used to draw out the knowledge from spatial databases. The knowledge secured is utilized to identify spatial and non-spatial data and their relation is inspected from [2,3] and the further information is analyzed from the same. The spatial databases contain information that represents the spatial data [4]. Spatial data gives insights about objects that can be represented numerically in geographic coordinate system.

Spatial data is utilized for mapping and is represented through coordinate system and furthermore utilizing distinctive topologies.

From spatial databases, drawn out the knowledge is used by spatial data mining. To identify spatial and non-spatial data, knowledge secured is utilized and their relation is inspected from [2, 3] and the further information is analyzed from the same. The spatial databases contain information that represents the spatial data [4]. Spatial data gives insights about objects that can be represented numerically in geographic coordinate system. Spatial data is utilized for mapping and is represented through coordinate system and furthermore utilizing distinctive topologies. From a large number of spatial databases, meaningful and useful results are discovered from many techniques and algorithms [5]. For spatial data, one of the major data mining methods is clustering. Grouping a set of spatial objects into groups called clusters and this process is known as spatial clustering [6]. A high degree of similarity is shown by objects inside a cluster. The clustering is performed by number of different methods, but partitional clustering, hierarchical clustering, and locality-based clustering are the three main divisions. A partition of the data is developed by partitional clustering such that objects in a cluster are not similar than objects in other clusters, but the object in a cluster are similar to each other [7]. Forms of partitional clustering are k-means and k-medoid methods. A sequence of partitioning operations are performed by hierarchical clustering.These can be done bottom-up, performing repeated amalgamations of groups of data until some pre-defined threshold is reached, or top-down, recursively dividing the data until some pre-defined threshold is reached. Document and text are frequently analyzed by hierarchical clustering. Hierarchical clustering is grid-based clustering. Based on local relationships, objects are grouped by locality-based clustering algorithms and therefore scanning of the entire database can be in one pass. While others assume a random distribution, a density-based algorithms are some locality based algorithms.

The proposed method [8] identifies clusters of different shapes, sizes, and densities. It requires only three parameters; these parameters take only integer values. So, it is easy to determine. The experimental results demonstrate the superior of the proposed method in identifying varied density clusters.

To identify significant areas containing spatial objects like object with spatial attributes, clustering technologies are widely required in many geography-related problems. Geographic non-overlapping constraint should be satisfied by the resultant geographic areas at most of times. That is, the overlapping of areas with other areas should not be allowed. For solve the overlapping problem, many approaches have been proposed. However, the non-overlapping condition cannot be guaranteed by many clustering methods.

Objectives of the Research

- To select an optimal centroid values and k range using Adaptive Adjustment Factor based Glowworm Swarm Optimization (AAFGSO) algorithms for achieving higher clustering accuracy.
- Weighted convolutional Neural Networks (Weighted CNN) classifier to reduce time complexity.
- To satisfy the non-overlapping condition using Enhanced Penalized Spatial Distance (EPSD) Measure.
- To verify the performance of clustering algorithms using spatial data.

The organization of the paper is as follows: An overview of spatial cluster algorithms and the underlying concepts are presented in section II. Non-Overlapping Constraint based Optimized K-Means with Density and Distance-based clustering (NOC-OKMDD) approach in detail is presented in section III. After that section IV presents a discussion on the methods. Finally, a brief conclusion is mentioned by section V.

2. Earlier Works

Based on proximity graphs, acut-Edge algorithm for Spatial Clustering (CutESC) is designed by Aksac et al[9]. When a cut-edge value for the edge's endpoints is below a threshold, edges are removed by the CutESC algorithm. Based on its neighborhood, statistical features and spatial distribution of data is used to calculate the cut-edge value. While automatically discovering clusters with non-uniform densities, arbitrary shapes, and outliers, the algorithm works without any prior information and preliminary parameter.

For geochemical anomaly detection, a new approach is designed for Integration of auto-encoder network with density-based spatial clustering by Zhang [10]. Noise sample (e.g., geochemically anomalous) which is differing from core samples (e.g., geochemically background samples) as anomalies can be regarded by the density-based spatial clustering application with noise (DBSCAN) and that is compared with deep auto-encoder network. Therefore, the DBSCAN is used to detect the noise samples representing geochemical anomalies and this DBSCAN clusters the learned representations from the code layer in the auto-encoder network. A novel clustering method, called SCDOT (Spatial Clustering with Density-Ordered Tree) is designed by Cheng et al [11]. A Density-Ordered Tree is formed by projecting a dataset. With a box-plot method, the data is portioned into several relatively small sub-clusters by SCDOT. A sub-cluster is merged repeatedly to find the genuine clusters and this method is called heuristic method. To determine input parameters automatically, an iteration strategy is used. Moreover, Cluster center and noise is identified by providing an innovative way. Based on topological features computed over the persistence diagram, data clustering approach is presented by Pereiaet al [7] and the theory of persistent homology is used to estimate it. The topological

properties such as Betti numbers, i.e., *n*-dimensional holes count in the discrete data space is indicated by this feature. The clustering of time series is enabled by the main contribution of the designed system and their attractor characterizes their similar recurrent behavior in phase space and similar scale-invariant spatial distributions are in spatial data. The point-to-point dissimilarity measures such as Euclidean distance or elastic measures relied by them and this information is ignored by it [12].

Bappee et al [13] to assemble an AI model for wrongdoing forecast utilizing geospatial highlights for various classifications of wrongdoing. The turnaround geocoding strategy is applied to recover Open Street Map (OSM) spatial information. This structured work likewise proposes discovering hot points extricated from wrongdoing hotspots region found by Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). A spatial separation highlight is then processed dependent on the situation of various hot points for different sorts of wrongdoing and this worth is utilized as a component for classifiers from the outcomes watched a critical presentation improvement in wrongdoing forecast utilizing the new created spatial highlights. Muhammad Zulqarnain [14] Predicting Financial Prices of Stock Market using Recurrent Convolutional Neural Networks. Ahmed Fahim [15] proposes a solution for these two problems together; by adding a preprocess step to get the expected number of clusters in data and better initial centers.

3. Proposed Methodology

In this proposed phase, developing spatial data mining field for multidimensional data is explored. Dataset incorporates all x and y directions with class label. It is utilized to separate examples from spatial information. Figure 1 shows proposed work's flow diagram. Fisrt step to select an optimal centroid values and k range using Adaptive Adjustment Factor based Glowworm Swarm Optimization (AAFGSO) algorithms for achieving higher clustering accuracy.Second,Weighted convolutional Neural Networks (Weighted CNN) classifier to reduce time complexity. Third, to satisfy the non overlapping condition using Enhanced Penalized Spatial Distance (EPSD) Measure. To verify the performance of clustering algorithms using spatial data.

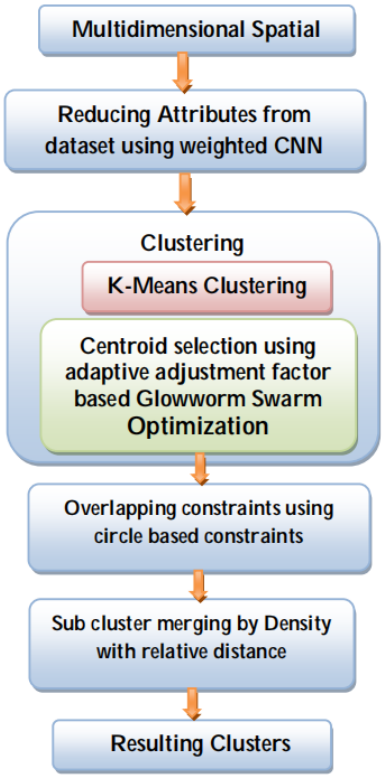

Fig.1. Flow diagram of the proposed system

3.1. Attribute Reduction using Weighted Convolutional Neural Networks (Weighted CNN)

There exists an output, input and several hidden layers in convolutional neural network [16,17,18]. Series of convolutional layers are in typical hidden layer, where they convolve using multiplication or dot product. Input to network is a multidimensional spatial dataset. The input layer which takes the spatial dataset as the input and the output layer provide the dataset reduction (attribute reduction).

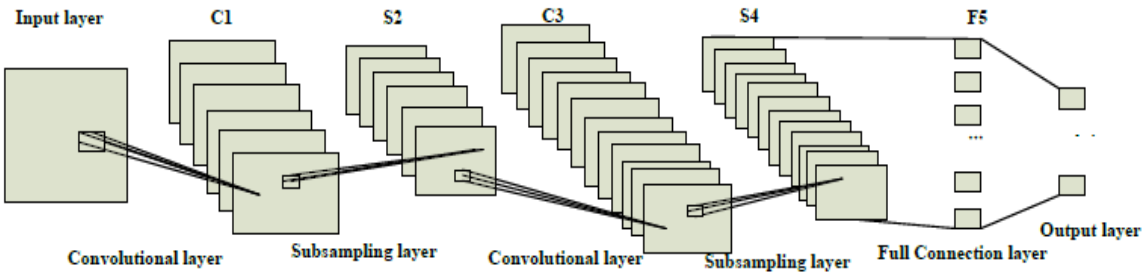

Fig.2. A schematic presentation of CNN

It works on data to compresses and makes smooth data. Max-layer selects the maximum value of the receptive field and produces data invariant to small translational changes. Consequently," generate three CNNs layers to manage various data prediction due to their variances in sizes. When apply sub sampling layers and for final output using fully connected layer. This structure of CNN enables the model to acquire filters that is capable to identify particular features in the input data. Due to the layered architecture of CNNs, they able to perform better on noisy data, by removal noise in each subsequently layer and capturing only the useful information [14].

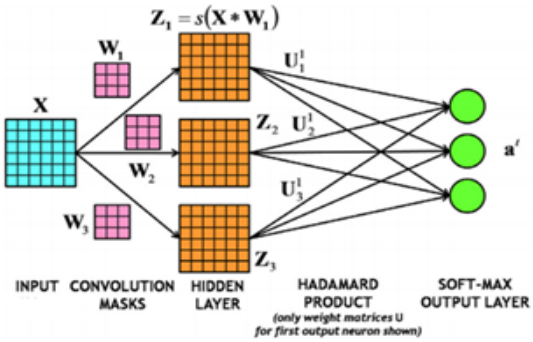

Fig.3. Weighted Convolutional Neural Network (Weighted CNN)

The info spatial dataset X convolves with the 3 covers W1, W2, and W3. These covers go about as open fields in the spatial dataset. The subsequent quality decrease through the concealed neurons is performed with strategic sigmoid actuations Then the weighted CNN processes component savvy Hadamard items between the shrouded neuron enactment networks Z1, Z2, and Z3 with weight frameworks $U_j^k$where j = 1, 2, 3 and k = 1, 2, 3.

Assume X as a input multidimensional data with $M_X \times N_X$size, where, M_X and N_X are positive values.2D filters $W_1 \ldots W_J$ are each of size $M_W \times N_W$. Matrix C_j is produced by convolving X with filter $W_j$and is given by,

$$C_j = X * W_j \tag{1}$$

Where,2D convolution is represented using.

Size of a2D data matrix C_j is $(M_X + M_W - 1) \times (N_X + N_Y - 1)$ and has $(m, n)^{th}$ entry

$$C_j(m,n) = \sum_{a=1}^{M_W} \sum_{b=1}^{N_W} X(a-m, b-n) W_j(a,b) \tag{2}$$

In above double sum, in order to define X in all points, it is padded with zeros. Through a logistic sigmoid functions s, J matrices $C_1, \ldots, C_J$ is passed element wise for producing activations of hidden neurons Z_j,

$$Z_j(m,n) = s\left(C_j(m,n)\right) = \frac{1}{1+\exp\left(-C_j(m,n)\right)} \tag{3}$$

If there exist a K output neurons in network, j^{th} hidden neuron matrix Z_j is multiplied element wise using $M_X + M_W - 1) \times (N_X + N_Y - 1)$ weight matrix U_j^k. At k of k^{th} output neuron, Gibbs or softmax activation is given as the ration of,

$$a_k^t = \frac{\exp\left(\sum_{j=1}^{J} e^T Z_j \odot U_j^k e\right)}{\sum_{k_1=1}^{K} \exp\left(\sum_{j=1}^{J} e^T Z_j \odot U_j^{k_1} e\right)} \tag{4}$$

Where, between two matrices, elementwise Hadamard product is represented as $\odot$, vector with all 1 is represented as e with a length ($M_X + M_W - 1$) ($N_X + N_W - 1$). Output and hidden neurons connection weight is represented using JK matrices U_j^k (j = 1, . . .Jand k = 1, . . ., K). Reduced the attributes according to the values of weight.

3.2. Non overlapping constraint based Optimized K-means with density and distance-based clustering (OKMDDC)

In this proposed research work, the reduced spatial dataset partitioned into sub clusters using K-means algorithm. However, it has issue with selection of number of clusters. so that cannot lead to the fruitful result. To solve this problem the proposed system designed an Adaptive Adjustment Factor Based Glowworm Swarm Optimization (AAFGSO) for k values selection.

Notion of clusters

In *d*-dimensional space R^d, *n* points set is represented as X = x_1, x_2… x_n, set of *K* sub clusters are represented as *S* =s_1, s_2,.. $s_K$and they are formed by splitting X using K-means, number of elements in s_k is represented as $|s_k|$, $D_k$is a $|s_k| \times K$ matrix is formed using K-means, whose entries having distance from every point of $s_k$to every *K* centroid. Distance between $s_i$ands_j is represented as d_{ij}, cutoff distance is represented as d_c. In X, genuine cluster set is represented as *C*= $c_1, c_2, .., c_G$.

Definition 1: Minimum distance between $s_i$points and $s_j$centroids is not always same as the distance between$s_j$points and s_i centroids. Thus, distance between $s_i$and $s_j$is computed using mean of these two values:

$$d_{ij}=d_{ji}=\frac{1}{2}\left[\min_{j}(D_i.j) + \min_{i}(D_j.i)\right] \tag{5}$$

Where, $D_{i:j}$is the j^{th} column of D_i and $D_{j:i}$isi^{th} column of D_j.

Instead of distance between sub cluster centroids, distance expressed in (1) is utilized in this work. More information regarding closeness between two sub clusters is given by the former definition and it is used for computing non-convex shape clusters.

Definition 2: *S*ub cluster *I's* local density $\rho_i$is defined by rewriting exponential kernel as,

$$\rho_i = \sum_{j=1}^{K} |s_k| \exp\left(\frac{d_{ij}^2}{d_c^2}\right) \tag{6}$$

Definition 3: Between subclusters *i* and other sub clusters having higher density, minimum distance δ is measured as,

$$\delta_i = \min_{j:\rho_j > \rho_i}(d_{ij}) \tag{7}$$

For the subcluster with highest density, $\delta_i = \max_j (d_{ij})$. The neighbor of sub cluster *i*(exception of sub cluster having highest density) is expressed as,

$$\text{Ne(i)}=\arg\min_{j:\rho_j > \rho_i}(d_{ij}) \tag{8}$$

A. Adaptive Adjustment Factor Based Glowworm Swarm Optimization

In this proposed research work, an Adaptive Adjustment Factor Based Glowworm Swarm Optimization algorithm (AAFGSO) is utilized for selecting optimal k values.

In object functions search space, randomly distributed glowworms swarm in GSO [19-20]. In glowworm algorithm, luminescence quantity which is termed as luciferin is carried by agents along with them. Neighboring glowworms brighter glow attracts, every glowworm. Neighboring glowworms are identified using glowworms in the position within current local-decision domain of glowworm. Current location fitness of glowworms is related to its luciferin intensity. In search space, glowworm location is made better, if it has high value of luciferin intensity. In every iteration, there will be a change in glowworms position and update of luciferin value. Every iteration has, luciferin update phase and transition rule based movement phase. There are four stages in GSO algorithms; they are Glowworms initial distribution, update of luciferin, movement and update of neighborhood range.

a. The initial distribution of glowworms phase

Glowworms initial distribution phase is also termed as initialization phase. In object functions search space, glowworms should be distributed randomly, which is a major objective, and it is distributed like a data point in dataset. Same intensity luciferin is carried by these glowworms and their decision domain r_0 are also same.

b. Luciferin-update phase

Based current fitness value of every glowworm, every glowworm is updated using luciferin in luciferin update phase. Accuracy of clustering is defined using fitness value. The current value of lucifering requires to be added to its current value of fitness based on its value in the previous moment. Value of luciferin is volatilized as time passage is required to be removed. Formula of specific luciferin update is expressed as,

$$l_i(t+1) = (1-\rho)l_i(t) + \gamma f(x_i(t)) \tag{9}$$

$f(x_i(t))$ -clustering accuracy

Where, at time t, level of luciferin associated with glowworm is represented as $l_i(t)$, luciferin decay constant is represented as q and it lies between 0 to 1, luciferin enhancement constant is represented as γ and at time t at agent i's location, objective function value is represented $f(x_i(t))$.

c. Movement phase

Probabilistic mechanism is used in every glowworm for making a deciding to move towards a neighbor having high luciferin value than its own value in movement phase. Glowworms having high brightness will attracts neighboring glowworms. At time t, glowworm i's neighboring set is computed as [21],

$$N_i(t) = \{j: ||x_j(t) - x_i(t)|| < r_d^j(t);\ \ l_i(t) < l_j(t)\} \tag{10}$$

Where, Euclidean norm operator is represented as ||.||, at time t, variable neighborhood range associated with glowworm i is represented as $r_d^j(t)$ and it is bounded in sensor range $r_s (0 < r_d^i(t) \leq r_s)$. For every glowworm i, moving probability towards neighbor j is represented as,

$$p_{ij}(t) = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)} \tag{11}$$

Where, j ∈ $N_i(t)$: assume glowworm j is selected by glowworm i with probability $p_{ij}(t)$ and it is expressed in (7). Movement of glowworm is expressed as,

$$x_i(t+1) = x_i(t) + s\left(\frac{x_j(t) - x_i(t)}{||x_j(t) - x_i(t)||}\right) \tag{12}$$

Where, in d-dimensional real space R^d, in at time t, glowworm i's location is represented as $x_i(t) \in R^d$ and step size is represented as s and it will be > 0.

d. Neighborhood range update rule

Every agent i is associated with a neighborhood having dynamic radial range $r_d^j(t)$in nature. Every glowworm's initial neighbourhood range is represented r_0. Every glowworm's rule is updated for updating range of neighborhood adaptively. Applied the following rule for the same.

$$r_d^j(t+1) = \min\{r_s, \max\{0, r_d^j(t) + \beta(n_t - |N_i(t)|)\}\} \tag{13}$$

Where, constant parameter is represented as β, parameter controlling neighbours count is represented as n_t.

In standard GSO, large jump is covered by every glowworm, if the value of fixed step-size is high. Therefore, in update, optimum solution may be missed due to this fast movement of glowworms. Glowworm will start to oscillate, if the distance between glowworm and its best neighbor is less than a value given in (12). Convergence rate is reduced, if there is a decrease in step-size. Finding the proper step size value is a difficult problem. In this research work, in order to overcome this difficulty, in every iteration, step size is varied instead of having fixed value.

Based on expression (14), enhanced the search step size for improving glowworm's search accuracy. In this enhanced expression, added an adaptive adjustment factor α. This is done to provide ability to algorithm for making new space of solution around good solution for enhancing the ability in fine search.

$$s(t) = \alpha \times stepa_0 \quad (14)$$

Where, search step size's initial value is represented as $stepa_0$, adaptive adjustment factor is represented as α and it lies between 0 to 1.

Adaptively computed adaptive adjustment factor α as,

$$\alpha = 1 - \left(\left(\frac{H-1}{H_{max}}\right)^2\right) \quad (15)$$

Where,
H - Number of searches
H_{max} − Maximum search

Algorithm 1: Adaptive Adjustment Factor -Glowworm Swarm Optimization (AAF-GSO)

1. Dimensions count is set as m
2. Initialize glowworms count (data points)
3. Set the generation G = 1
4. Assume step size as s
5. In search space, uniformly as well as randomly distribute glowworms (data points)
6. While (G < max generation) do
7. for every glowworm (data points) i do
8. Based on expression (9), update luciferin
9. Based on expression (10), set of neighbors are confirmed
10. Based on expression (11), movement probability is computed
11. Based on expression (12), glowworm i moves toward j; for every glowworm, next position x(t+1) and next decision radius $r_d(t+1)$ are computed and adaptive adjustment factor which is expressed in (14) is used for renewing next step-size.
12. Based on expression (13), neighborhood range is updated
13. End for
14. End while

The K-means clustering is a widespread partitioning technique which has large number of applications. It is an iterative method to find the cluster centroids for each iteration to get the closest seed. The proposed method starts with the generation of optimum cluster centroids using AAFGSO in the first phase.

B. Overlapping constraints using Enhanced Penalized Spatial Distance (EPSD)

In numerous topography related issues, grouping innovations are generally required to distinguish huge regions containing spatial items. At the greater part of times, the resultant geographic territories ought to fulfill the geographic non-covering requirement. That is, the zones ought not be covered with different regions. A few reasons can create covering bunches: there may be commotion in the information, the highlights may not catch all the important data to obviously isolate groups, or the cover might be natural to the procedures that delivered the information. The many grouping strategies cannot ensure this condition. To take care of this issue the proposed framework planned an Enhanced Penalized Spatial Distance (EPSD) Measure, and it is sealed to fulfill the condition, which can ensure the imperative. The EPSD accomplishes this by well altering the spatial separation between two focuses as indicated by the spatial trait esteems between them. On the off chance, that the spatial quality qualities between two focuses change bigger, the EPSD will be punished (expanded) bigger. In the event that else, the separation will be punished less.

Considering the spatial trait as the multi measurement and utilizing Manhattan Distance, the above condition will not be kept if spatial quality is prevailing. On one line, three spatial points are represented asa∗, b∗, c∗and their spatial values are represented asa, b, c. b∗is in middle of a∗c∗, but Manhattan Distance are

D(a, b) > D(a, c). Thus, b∗may be in a various cluster from a∗andc∗.

The proposed inquire about present another separation measure named Enhanced Penalized Spatial Distance (EPSD), which well modifies the spatial separation between two focuses by the difference in the spatial property estimations between them. On the off chance, that the non-spatial characteristic qualities between two focuses change bigger, the EPSD will be punished (expanded) bigger. On the off chance that else, the separation will be punished less. The separation measure has the character: if *a*, *b*, and *c* are on one line in spatial domain, and *b* is in the middle of *ac*, then $D_{EPSD}(ac) = D_{EPSD}(ab) + D_{EPSD}(bc)$. Because $D_{EPSD}(bc) > 0$, $D_{EPSD}(ac) > {}_{DEPSD}(ab)$ can be always kept. So that this distance measure can satisfy above condition. Entire process has two steps like finding value series between two points in sub clusters and Adjust Spatial Distance

Suppose value series is $V_{ab} = \{S(g0), S(g1), S(g2), \cdots, S(gn)\} = \{s0, s1, s2, \cdots, s3\}$, where $s_i = S(g_i)$. $g_0$is start point,

and $g_n$is end point. Considering *i*-th value and (*i*+ 1)-th value, if difference between $s_{i+1}$and $s_0$is larger than $s_i - s_0$, distance will be penalized more. If else, distance will be penalized less. By starting from first value in $V_{ab}$and processing each value one by one, total PSD can be formulated in Equation

$$D_{EPSD}(a,b)=\sum_{s_i \in V_{ab}}\{D_{geo}(g_{i+1}, g_i) * p\,(s_{i+1}, s_i)*k_d \quad (16)$$

Where, control factor is represented as k_d and it describes the influence of spatial attribute on distance. Large influence is possessed by large value of k_d.

Every spatial attribute total change is used for adjusting spatial distance of every sub cluster. With Q attributes and $k_q$as a weight for *p*-th attribute, extended the expression (16) as,

$$D_{EPSD}(a,b)=\sum_{g_i \in P_{ab}}\{D_{geo}(g_{i+1}, g_i) * \sum_{q \in Q}\{k_q * p\left(S_q(g_{i+1}), S_q(g_i)\right)\}\} \quad (17)$$

C. Density based clustering with relative distance

The non-covering sub bunches are converged with the assistance of Density based grouping with relative separation approach. In a thickness based bunching calculation, D is dataset, Eps is the grouping range, and MinPts are the base number of articles. The neighboring information point inside the range of Eps of specific point in the dataset is the Eps-neighborhood of that point [22].

The test point is named as the Core point if the Eps-neighborhood of a test point including a base number of neighbors MinPts; at that point is viewed as a center point. Fringe point is resolved if the Eps-neighborhood of a test point comprehensive of least neighbors, MinPts, and distinguished to share its Eps-neighborhood with at least recognized center point. Straightforwardly thickness reachable: A point p is regarded as legitimately thickness reachable from another point q if p is inside Eps neighborhood of q and q is recognized as a center point as of Fig. 3.

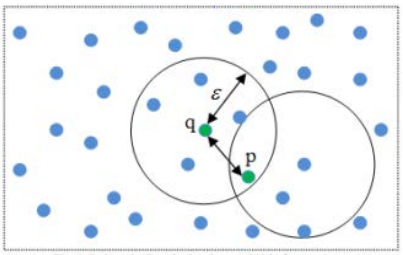

Fig.3. Point p is directly density-reachable from point q

Density reachable: In a dataset D, based on Minpts and Eps, point p is deemed as density reachable to q, if there is a data point chainp1, p2, ….pn, p1=q and pn = p, so that pi + is directly density-reachable from pi. Fig.4 demonstrates the density reachable definition.

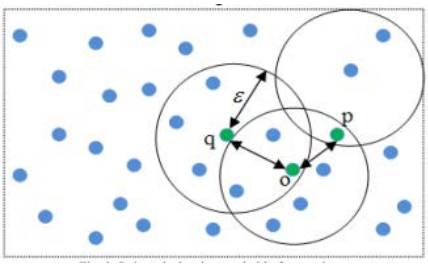

Fig.4. Point p is density-reachable from point q

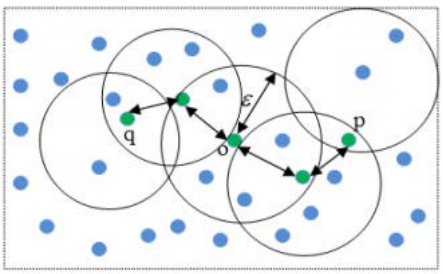

Fig.5. Point p is density connected to point q

Density connected: In a dataset D, based on Minpts and Eps, point p is deemed as density connected to q, if there is a data point o ∈ D, so that, based on Minpts and Eps, point p and q are density-reachable from point o. Fig.5 demonstrates the density connected definition.

In dataset, every test point's Eps-neighborhood is checked in DBSCAN for dataset clustering. New cluster with core point p is created, if minimum number of neighbors, MinPts is less than test data point's Eps-neighborhood. From this core point, density reachable points are re-accumulated in this algorithm directly and are merged with density reachable clusters. Terminated this process while null new point to any group. The excluded points of cluster are represented as outliers or noise. Subsequent steps are steps of DBSCAN:

- Initiate with a random point p
- Retrieves every point which are density-reachable from p for distance, Eps (ε), MinPts.
- If p is identified as a major point, cluster w.r.t. Eps(ε) and MinPtsis considered in this method.
- If p is identified as a boundary point excluding density-reachable from p, DBSCAN moves to next database data point.

Algorithm 4: Non-overlapping constraint based OKMDD algorithm
Input: dataset X, sub clusters count K, clusters count G, cutoffdistanced_c.
Output: Cluster C

1. Initialize Multidimensional spatial dataset
2. The number of sub clusters K are computed using Adaptive Adjustment Factor -Glowworm Swarm Optimization (AAF-GSO) optimization
3. Use K-means for partitioning dataset X into K sub clusters
4. Compute the distance between each pair of sub clusters d_{ij}
5. Compute ρ of each sub cluster by (6)
6. Compute δ of each sub cluster by (7), and get its neighbour by (8)
7. Use the decision graph to select G cores and (if need) noise sub clusters.
8. Create G clusters for G cores
9. Compute non overlapping constraints using Enhanced Penalized Spatial Distance (EPSD)
10. If sub cluster i(i=1, 2,..,K) is not a core or not belongs to noise
11. Assign it to the cluster its neighbor belongs to Else merge them End if

Label each point according to the cluster label of its belonging sub cluster, and get the final clusters C=$\{c_1, c_2, \ldots, c_G\}$.

4. Experimental Results

The existing and proposed clustering algorithms are implemented in the MATLAB. The data are collected from http://www2.cs.uh.edu/~ml_kdd/restored/Complex&Diamond/Complex9.txt. Evaluate the designed method on the above dataset. Here we give a comparison between various clustering models namely, K-means, OKMDD, OKMDD

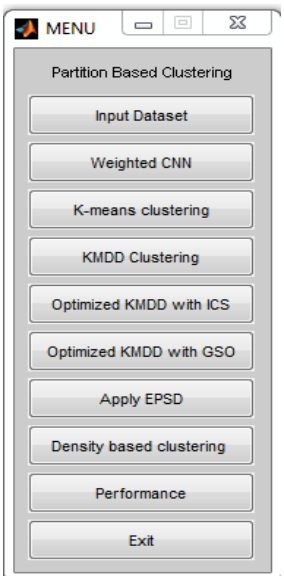

Fig.6. Main menu

with Improved Cuckoo Search (ICS) and OKMDD with Adaptive Adjustment Factor-Glowworm Swarm Optimization (AAF-GSO) algorithm with respect to types of matrices namely adjusted rand index, rand index, mirkins index and Hubert's index.

The simulation results of the proposed research technique are illustrated in the Fig.6 to 8, Fig. 6 illustrates the main menu function of the proposed technical work.

The optimal clustering is shown in Fig. 8 and 9. In this proposed research work, non-overlapping constraint based Optimized K-means with density and distance-based clustering (OKMDDC) is performed. In order to select an optimal k value, Adaptive Adjustment Factor Based Glowworm Swarm Optimization algorithm (AAFGSO) is utilized.

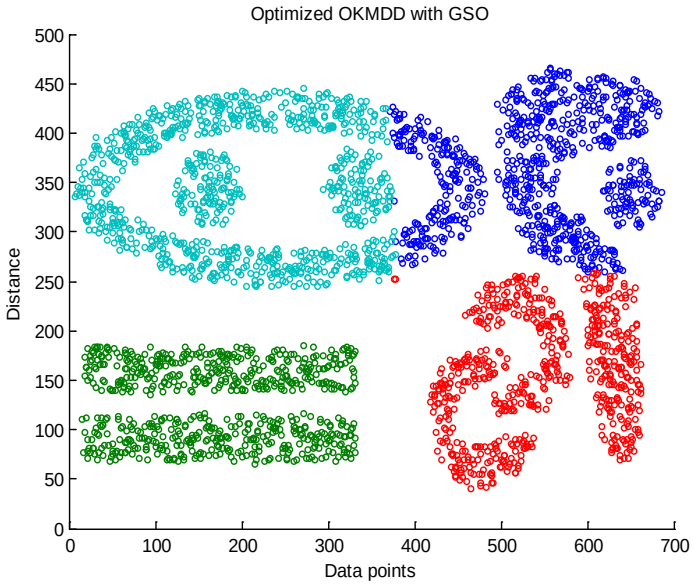

Fig.7. Optimized OKMDD with GSO

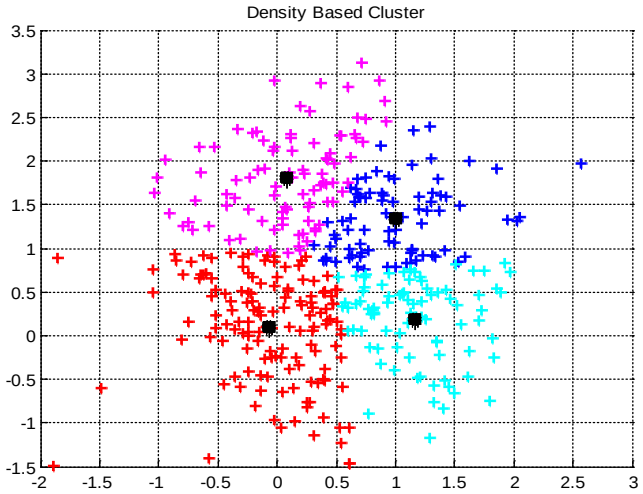

Fig.8. Density Based Clustering

4.1. Performance evaluation

From the Fig.9, the performance of proposed OKMDD with AAF-GSO algorithm is compared with the existing K-means, KMDD, and OKMDD with ICS algorithm in terms of adjusted rand index. In x-axis, clustering techniques are represented and adjusted rand index is represented in y-axis. From experimental results, proposed system has 81 as an adjusted rand index whereas other methods such as K-means, KMDD, OKMDD with ICS has 19, 46 and 63 respectively.

Rand index of the proposed OKMDD with AAF-GSO algorithm is compared with the existing K-means, KMDD, and OKMDD with ICS algorithm in terms of rand index. In x-axis, clustering techniques are represented, and rand index is represented iny-axis. Investigational outcome shows that the designed system attains rand index of100 when the other method such as k-means, KMDD, OKMDD with ICS achieves 39, 66 and 83 respectively.

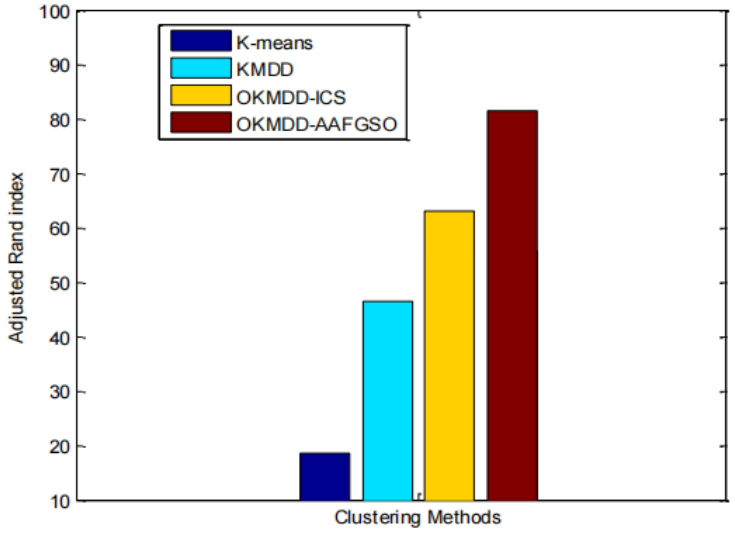

Fig.9. Adjusted rand index comparison

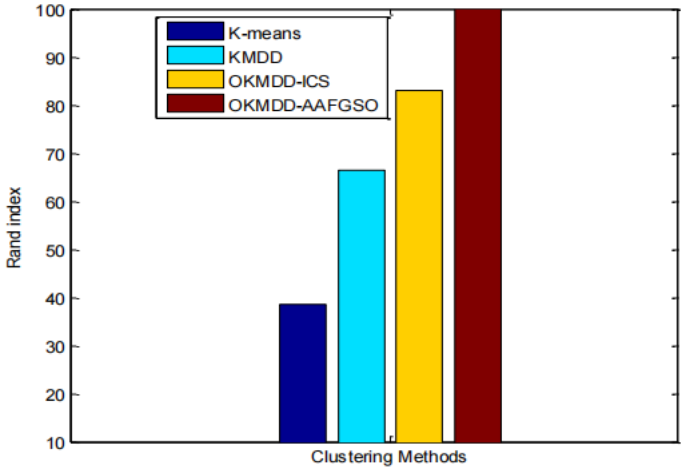

Fig.10. Rand index comparison

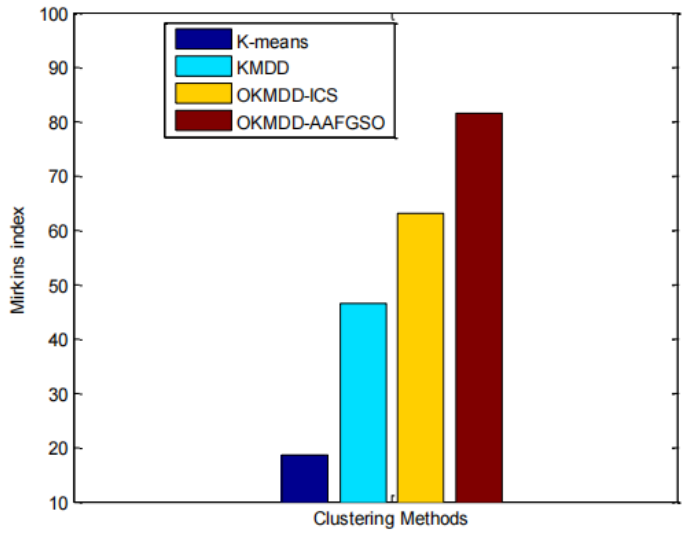

Fig.11. Mirkins index comparison

The Mirkins index of the proposed and existing methods is shown in Fig.11. In x-axis, clustering techniques are represented and Mirkins index is represented in y-axis. The investigational outcome shows that the designed system has 81 as a Mirkins index when the other methods such as k-means, KMDD, OKMDD with ICS achieves 19, 46 and 63 respectively.

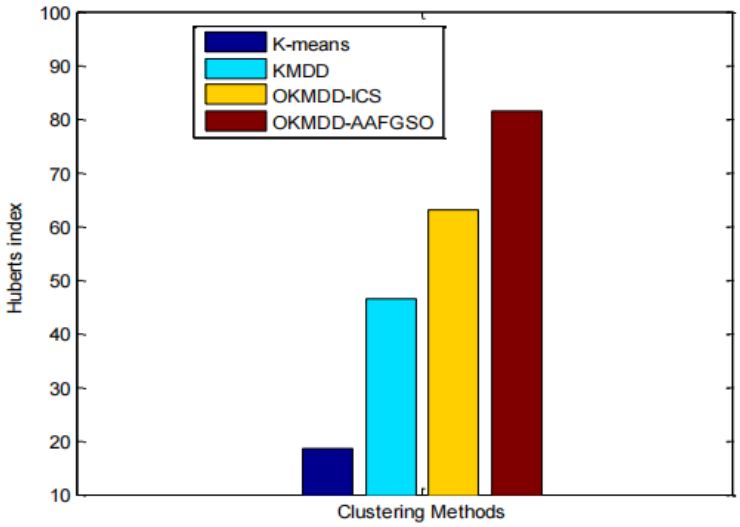

Fig.12. Huberts index comparison

Fig.12 shows the performance of proposed OKMDD with AAF-GSO algorithm and existing K-means, KMDD, OKMDD with ICS algorithms in terms of Hubert index. In x-axis, clustering techniques are represented, and Hubert index is represented in y-axis. From experimental results, the proposed system has 81 as Huberts's index whereas other methods such as K-means, KMDD, OKMDD with ICS has 19, 46 and 63 respectively.

5. Conclusion

In this paper we have introduced a very simple idea to discover clusters with various shapes, sizes, densities and non-overlapping constraints. The proposed research work designed a Non-Overlapping Constraint based Optimized K-Means with Density and Distance-based Clustering (NOC-OKMDDC) method for multidimensional spatial database. For reducing multi-dimensional spatial dataset attributes, weighted Convolutional Neural Networks (Weighted CNN) are performed. In Optimized K-Means with Density and Distance-based Clustering (OKMDD), the optimal number of sub clusters are done by using Adaptive Adjustment Factor based Glowworm Swarm Optimization algorithm (AAFGSO). It improves the clustering accuracy. In order to satisfy the non-overlapping constraint, the new Enhanced Spatial Distance Measure (EPSD) is utilized. In contrast with conventional agglomerative hierarchical clustering techniques, OKMDD is a first one, which employed a concept of distance and density for aggregating sub clusters.

This algorithm produces very good results as shown in Fig. 9,10,11,12. CNN models can further be improved with fine-tuning, and other machine learning algorithms to minimize the execution time. Furthermore, it will be remarkable to observe future work on implementing proposed model for further time series applications such as weather forecasting, earthquake prediction and signal processing.

References

[1] Arbind Kumar Singh and Manimannan, "Detecting Hot Spots on Crime Data Using Data Mining and Geographical Information System", International Journal of Statistika and Mathematika, ISSN: 2277- 2790, E-ISSN: 2249-8605, Volume 8, Issue 1, 2013 pp 05-09.

[2] Bendechache,Malika,andMTaharKechadi "Distributed clustering algorithm for spatial data mining", 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015.

[3] Parimala, M., Lopez, D., &Senthilkumar, N. C. (2011). A survey on density based clustering algorithms for mining large spatial databases. International Journal of Advanced Science and Technology, 31(1), 59-66.

[4] Mumtaz, K., &Duraiswamy, K. (2010). An analysis on density based clustering of multi-dimensional spatial data. Indian Journal of Computer Science and Engineering, 1(1), 8-12

[5] Barua, H. B., Das, D. K., &Sarmah, S. (2012). A density based clustering technique for large spatial data using polygon approach. TDCT, IOSR Journal of Computer Engineering (IOSRJCE) ISSN, 2278-0661

[6] Sharma, A., Gupta, R. K., & Tiwari, A. (2016). Improved Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data. Mathematical Problems in Engineering, 2016.

[7] Gupta, R. K., & Tiwari, A. (2015). Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data. Mathematical Problems in Engineering, 2015.

[8] Ahmed Fahim, "A Clustering Algorithm based on Local Density of Points", International Journal of Modern Education and Computer Science, Vol.9, No.12, pp. 9-16, 2017.

[9] Aksac, A., Özyer, T., &Alhajj, R. (2019). CutESC: Cutting Edge Spatial Clustering Technique based on Proximity Graphs. Pattern Recognition.

[10] Zhang, S., Xiao, K., Carranza, E. J. M., Yang, F., & Zhao, Z. (2019). Integration of auto-encoder network with density-based spatial clustering for geochemicalanomaly detection for mineral exploration. Computers & Geosciences.
[11] Cheng, Q., Lu, X., Liu, Z., Huang, J., & Cheng, G. (2016). Spatial clustering with density-ordered tree. Physica A: Statistical Mechanics and its Applications, 460, 188-200.
[12] Pereira, C. M., & de Mello, R. F. (2015). Persistent homology for time series and spatial data clustering. Expert Systems with Applications, 42(15-16), 6026-6038.
[13] Fateha Khanam Bappee, Amilcar Soares and Stan Matwin, “Predicting Crime Using Spatial Features”, March 2018.
[14] Muhammad Zulqarnain, Rozaida Ghazali, Muhammad Ghulam Ghouse, Yana Mazwin Mohmad Hassim, Irfan Javid, "Predicting Financial Prices of Stock Market using Recurrent Convolutional Neural Networks ", International Journal of Intelligent Systems and Applications, Vol.12, No.6, pp.21-32, 2020.
[15] Ahmed Fahim, "Finding the Number of Clusters in Data and Better Initial Centers for K-means Algorithm", International Journal of Intelligent Systems and Applications, Vol.12, No.6, pp.1-20, 2020.
[16] Vedaldi, A., &Lenc, K. (2015, October). Matconvnet: Convolutional neural networks for matlab. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 689-692). ACM.
[17] Niepert, M., Ahmed, M., &Kutzkov, K. (2016, June). Learning convolutional neural networks for graphs. In International conference on machine learning (pp. 2014-2023).
[18] Wang, M., Liu, B., &Foroosh, H. (2017). Factorized convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 545-553).
[19] Huang, Z., & Zhou, Y. (2011). Using glowworm swarm optimization algorithm for clustering analysis. Journal of Convergence Information Technology, 6(2), 78-85.
[20] Aljarah, I., & Ludwig, S. A. (2013, June). A new clustering approach based on glowworm swarm optimization. In 2013 IEEE Congress on Evolutionary Computation (pp. 2642-2649). IEEE.
[21] Oramus, P. (2010). Improvements to glowworm swarm optimization algorithm. Computer Science, 11, 7.
[22] Parimala, M., Lopez, D., &Senthilkumar, N. C. (2011). A survey on density based clustering algorithms for mining large spatial databases. International Journal of Advanced Science and Technology, 31(1), 59-66.

Authors’ Profiles

K Lakshmaiah, Research Scholar, received B.Tech Degree in Computer science and engineering, from Jawaharlal Nehru Technological University, Hyderabad, in 2002 and M.Tech Degree in Computer science and engineering, received from Sathayabhama University, Chennai, Tamilnadu, in 2007, and Ph.D pursuing from Jawaharlal Nehru Technological University Hyderabad, Hyderabad, Telangana.India.

Dr. S.Murali Krishna, received B.Tech Degree in Computer science and engineering from Sri Venkateswara University, Tirupati,Andhra Pradesh, in 2002 and M.Tech Degree in Computer science and engineering, received from Jawaharlal Nehru Technological University , Hyderabad, in 2005. He received Ph.D from Jawaharlal Nehru Technological University, Ananthapur, Ananthapuram, Andhra Pradesh, in 2011. He is a Professor and Head in Information Technology Department, Sri Venkateswara College Of Engineering, Tirupati, Andhra Pradesh, India. His research interests include Data mining, Text mining, machine learning distributed classification and clustering.

Dr. B.Eswara Reddy, received B.Tech Degree in Computer science and engineering from Sri Krishna Devaraya University, Ananthapur, Andhra Pradesh, in 1995 and M.Tech Degree in Computer science and engineering, received from Jawaharlal Nehru Technological University, Hyderabad, in 1999.He received Ph.D from Jawaharlal Nehru Technological University, Hyderabad, Andhra Pradesh, in 2008. He is a Professor in Computer Science and Engineering Department, Jawaharlal Nehru Technological University Ananthapur, Ananthapuramu, Andhra Pradesh, India. His research interests include Data mining, pattern recognition, Image processing.

How to cite this paper: K Laskhmaiah, S Murali Krishna, B Eswara Reddy, "An Optimized K-means with Density and Distance-Based Clustering Algorithm for Multidimensional Spatial Databases", International Journal of Computer Network and Information Security(IJCNIS), Vol.13, No.6, pp.70-82, 2021. DOI: 10.5815/ijcnis.2021.06.06