Density Based Clustering using Modified PSO based Neighbor Selection

K. Nafees Ahmed
Research Scholar, Dept of Computer Science
Jamal Mohamed College (Autonomous), Tiruchirappalli, India
nafeesjmc@gmail.com

Dr. T. Abdul Razak
Associate Professor, Dept of Computer Science
Jamal Mohamed College (Autonomous), Tiruchirappalli, India
abdul1964@gmail.com

***Abstract*—Density based clustering basically operates by associating related items contained in the sample space. The association is performed by maintaining maximum inter class similarity and minimum intra class similarity. However, the major downside of such approach is that it is time consuming in case of huge datasets. This paper proposes a metaheuristic based density clustering technique that utilizes a modified Particle Swarm Optimization (PSO) for fast and efficient neighbor selection. In this work, the PSO is integrated with simulated annealing to perform faster node selection and the distribution of catfish particles in the search space helps to avoid local optima to the maximum extent. Experiments were conducted with real-time spatial datasets and it was identified that the proposed clustering technique performs effectively in terms of both time and efficiency.**

***Keywords*-**DBSCAN; PSO; Catfish Particle; Simulated Annealing; Spatial Clustering.

I. INTRODUCTION

Clustering refers to grouping data in terms of some similarity measures. The grouping is performed such as to maintain minimum heterogeneity within group and maximum heterogeneity between groups. Clustering techniques are categorized as Density-based algorithms, Hierarchical algorithms, Partitioning algorithms, Combinational algorithms, Graph-based algorithms, Grid-based algorithms and Model-based algorithms. Each of these algorithms has their own distinct use cases. This work focuses on density based clustering techniques. Spatial data considered for the proposed approach is usually high dimensional, contains groups of varied shapes, sizes and densities, hence defining them on the basis of shapes or fixed number of groups is not possible. These tend to be the major reasons for selection of density based clustering. Further, identifying outliers is only possible by using a flexible algorithm.

The major use cases for density based clustering techniques include region planning, soil classification, social science, anomaly detection, biomedical image analysis and environmental quality evaluation. However, there are several issues or requirements existing in the current density based clustering techniques. Domain knowledge is mandatory, as the input parameters are user defined. However, when dealing with very large databases, this is not feasible. They are required to identify clusters of any shape, which may be non-convex, spherical, drawn-out, linear, elongated etc. Efficiency of these algorithms on large databases is questionable.

Density Based Spatial Clustering of Application with Noise (DBSCAN), developed by Ester et al. [1] is the base for algorithms that were developed for applications requiring density based grouping of data points. It operates on the basis of distance threshold and the density of neighboring points to integrate a point into an already defined cluster. The major limitations of DBSCAN are that the performance of the algorithm depends on the input parameters *minPts* and *maxThresh*. These points are to be provided as manual inputs; hence the accuracy of clusters depends on the efficiency of the inputs. Scalability is a major issue, leading to huge time consumption as the number of data points increase. Selection of the initial cluster points also plays a vital role in the accuracy of the clusters formed. Most of the available density based clustering algorithms are variants of DBSCAN, trying to overcome one or more of these limitations.

The rest of this paper is organized as follows: Section II provides the related works, Section III describes about the proposed methodology, Section IV presents the results and discussion. Section V concludes the paper with some future works.

II. RELATED WORKS

Density based clustering techniques are currently on the raise due to the huge scope of applications available under them. A variant of DBSCAN for operating on complex data was proposed by Mai et al. [2] called A-DBSCAN-XS, operates on complex data and was mainly designed to solve the issue of scalability. The algorithm operates by producing fast results; however, the results are near optimal. The near optimality levels of the results are enhanced by continuous pruning of the data. The major advantage of this technique is that the pruning phase can be stopped at any required time to obtain results. Though the results are intermediate, they are still suboptimal, hence eliminating the time constraint associated with the algorithm. Density based clustering techniques are limited to continuous data. In order to solve this issue, Azzalini et al. [3] proposed a density based clustering technique that operates on the non-continuous data effectively to provide appropriate groupings.

Even though density based clustering techniques are good at identifying outliers, they are also categorized as a cluster and are not eliminated from the dataset. An automatic outlier removal technique inspired by artificial immunity and density based clustering was proposed by Paul et al. [4]. The major advantage of this approach is that this technique claims to produce accurate clusters even with low inter-cluster distances. A feature selection technique operating on density based clustering was proposed by Sengupta et al. [5]. This technique operates on the basis of biological data. It uses a combination of density based clustering and feature dissimilarity measure to identify appropriate features in a biological data.

A reliable algorithm to extract shape specific features was presented by Luo et al. [6]. This technique also uses a combination of density based clustering and feature extraction techniques to identify dynamic shaped clusters in the data. It was mainly designed for uncertain data and operates on the density levels of data under consideration. Other variants of DBSCAN include FDBSCAN [7], which utilizes fuzzy theory to express data uncertainties, P-DBSCAN [8], that models uncertainties by associating the values with a probability density function (PDF). DBSCANEA [9] uses PDF to model the uncertainties and uses the maximum and minimum dissimilarity degrees as a measure. U-DBSCAN [10], GDBSCAN [11] also uses PDF and similarity measures to model uncertain data.

The above mentioned algorithms operate on raw data. The applicability of the clustering algorithms are limited to raw data, however when it comes to normalized data, due to the low interval levels, grouping becomes an issue. Another variant of spatial clustering used for hotspot analysis was presented in [12].

III. PROPOSED WORK

Clustering of appropriate data in the spatial domain has certain special requirements, hence making the conventional clustering techniques in-appropriate for usage. The requirements for spatial clustering includes dynamic cluster shape recognition, dynamic cluster count identification and appropriate outlier detection. However, these requirements tend to increase the time complexity of the cluster identification process.

The basic idea of a density based clustering algorithm is that it operates on the basis of the neighbor density of a node. A node is considered as a part of a cluster, if it has a group of neighbors >= *minPts*, satisfying the distance threshold, defined by *maxThresh*. This neighborhood based analysis acts as the basis for identifying varied shaped clusters without the need for providing the initial cluster count. However, certain input requirements become mandatory. The user is required to provide *minPts* and *maxThresh*, the minimum required points to qualify a node and the maximum threshold radius to be used for neighbor identification. Specifying these points becomes mandatory, as they vary considerably with respect to the datasets being used.

Particle Swarm Optimization is the metaheuristic based optimization technique used to solve optimization problems. The major issue in DBSCAN is that it analyzes each node for its containment, hence making the selection process time consuming. This process is replaced with PSO incorporated with catfish particles and Simulated Annealing. The density based spatial clustering technique proposed in this work is shown in the Fig. 1. It uses modified PSO based neighbor selection by integrating simulated annealing into its local search mechanism and also by the concept of catfish behavior into the selection process to reduce the problem of local optima, thereby speeding up the selection process.

The first phase of the proposed work is to analyze the data to identify these threshold points. The maxThresh is automatically identified by considering the 75th percentile of the distance limits, while the minimum required points are to be manually provided.

The initial cluster list remains empty and the node list is initialized with all the nodes in the search space. A random base node is selected and is passed to the modified PSO module to identify the neighbor nodes.

The particles are distributed on the base node and a random velocity is identified for each of these particles. This phase marks the beginning of the particle movement. The particles are distributed in the search space and the new velocity for the particle is calculated using the equation (1).

$$V_{i,d} \leftarrow \omega V_{i,d} + \varphi_p r_p \left(P_{i,d} - X_{i,d}\right) + \varphi_g r_g \left(g_d - X_{i,d}\right) \quad (1)$$

Where $r_p$and $r_g$are the random numbers, $P_{i,d}$ and g_d are the parameter best and the global best values, $X_{i,d}$ is the value current particle position, and the parameters ω ,φ_p, and $\varphi_g$are selected by the practitioner.

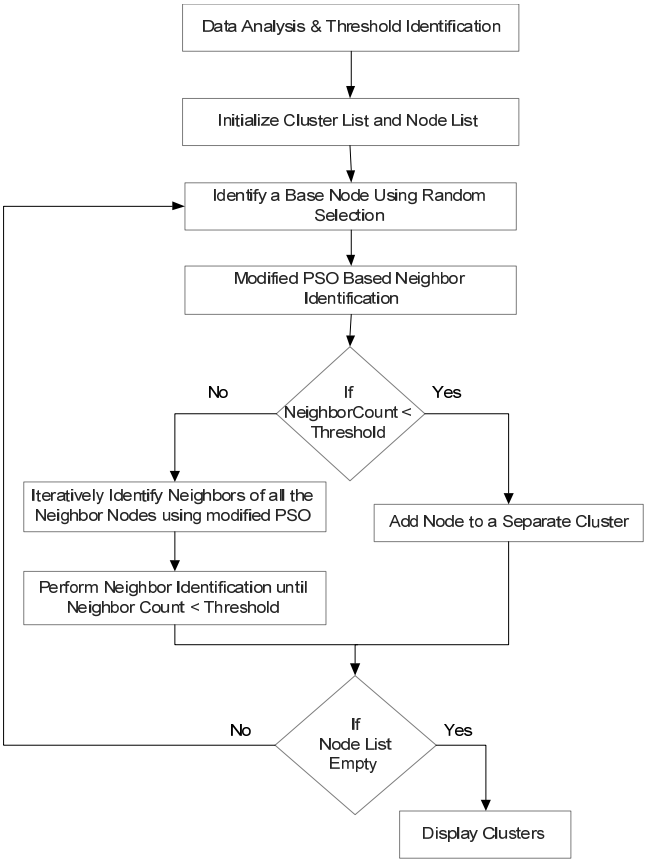

Figure 1. Density Based Spatial Clustering using Modified PSO based Neighbor Selection

Positions of the particles are altered using the new velocity component. The problem of clustering is discrete, whereas PSO operates on a continuous region, hence the resultant position is discretized and the particles are positioned on discrete nodes. Particle fitness is identified using simulated annealing and if the fitness is satisfied, if the node is not in the current neighbor list, it is added to the current neighbor list. Fitness of the solution is calculated as a function of the distance of the particle from the base node and its minimum number of neighbors.

Otherwise, the level of stagnation behavior is incremented. If the stagnation behavior reaches the catfish threshold (*CThreshold*) and if the current threshold hit is the initial hit, the particles are redistributed on the base node and the process is repeated, else the neighbor list is returned to the main module. These steps are diagrammatically shown in the Fig. 2.

The major reason for using a catfish particle is that PSO has high probabilities of getting struck in local optima. Once the *CThreshold* is reached, the particles are redistributed again and the entire process is performed again, hence eliminating the possibility of local optima.

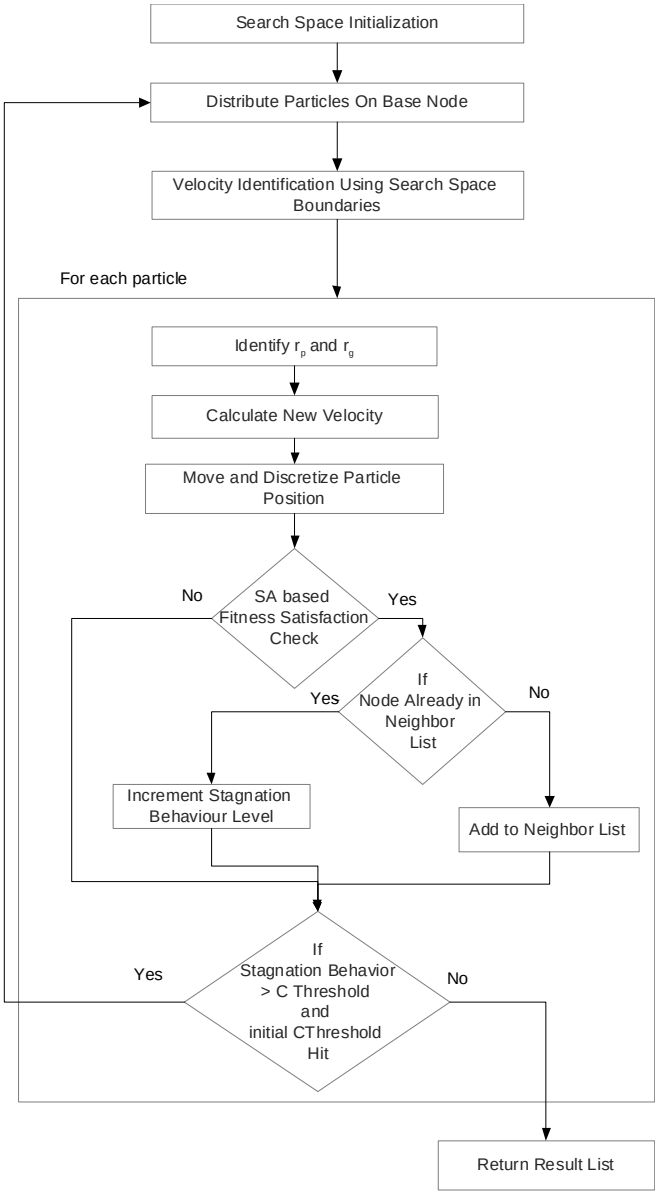

Figure 2. Modified PSO using Catfish particles and embedded Simulated Annealing

IV. RESULTS AND DISCUSSION

Analysis of the proposed modified PSO was performed by implementing it in C#.NET. Clustering specific datasets such as Iris, Banana, Quake and Forest were used for identifying the efficiency of the proposed algorithm.

Efficiencies were observed in terms of inter cluster distance, intra cluster radius (ICR), data density contained in each cluster and time.

Figures 3-6 exhibit the intra cluster radius of each of the datasets. The nodes with zero ICR are outliers. All other clusters contain at least *minPts* clusters. It could be observed that quake and banana exhibits high levels of outliers. Grouping of nodes are also effective in both the datasets. Forest dataset exhibits moderate clusters with low outliers, while Iris dataset exhibits five clusters, two clusters with high ICR, two with low ICR and one with moderate radius. These ICR values show that the clusters formed are not grouped just in terms of a defined radius, but in terms of the shape of the grouped points. This exhibits the highly efficient shape based aggregation efficiency of the proposed technique.

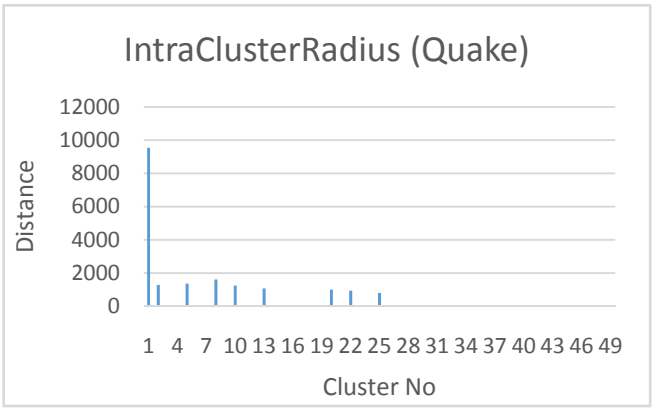

Figure 3. Intra Cluster Radius (Quake)

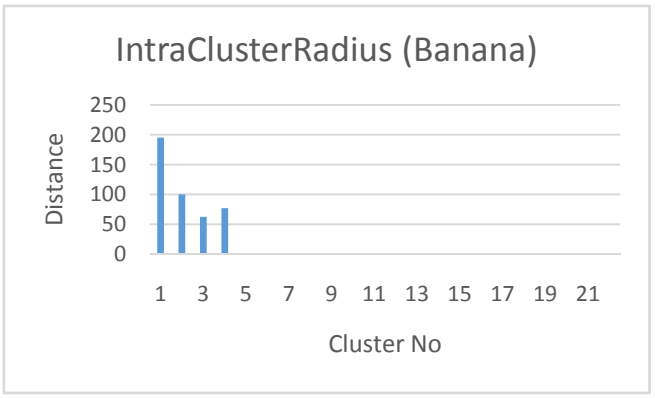

Figure 4. Intra Cluster Radius (Banana)

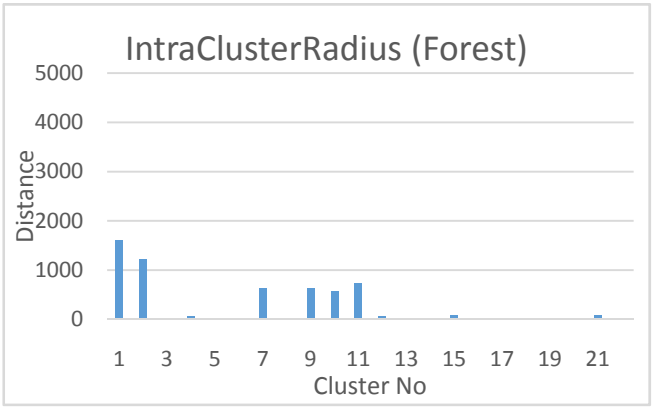

Figure 5. Intra Cluster Radius (Forest)

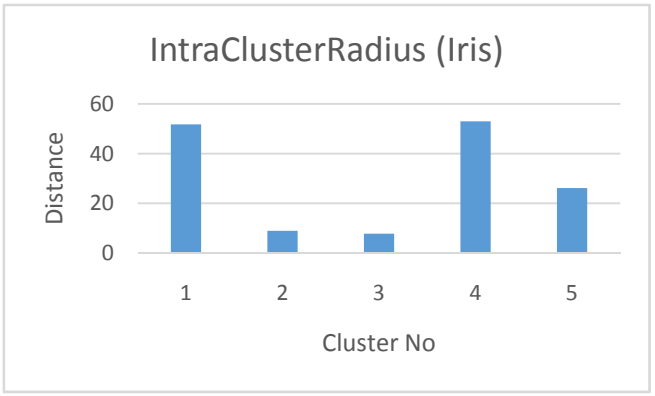

Figure 6. Intra Cluster Radius (Iris)

A Time comparison is carried out between the proposed modified PSO based clustering technique and multi-start PSO based clustering technique [13] is shown in Fig. 7.

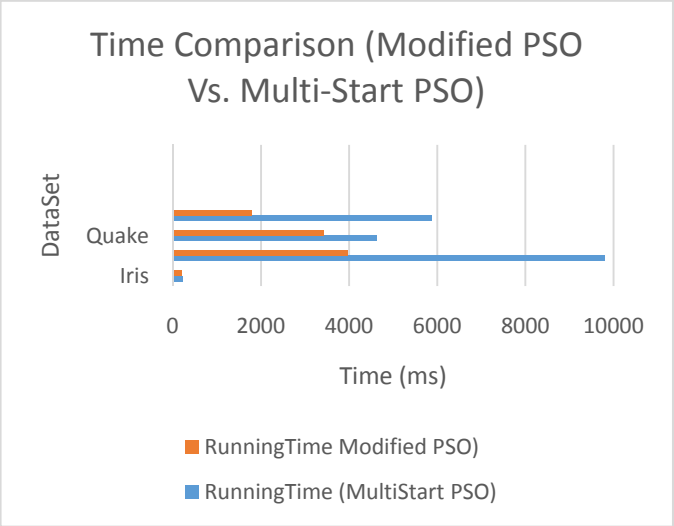

Figure 7. Time Comparison (Modified PSO vs. Multi-start PSO)

It could be observed that the proposed modified PSO based clustering technique exhibits 50% better computational time compared to the multi-start PSO.

Cluster density of each cluster corresponding to Quake, Banana, Forest and Iris are presented in Figures 8-11. It could be observed from the Quake, Banana and Forest datasets that several clusters of moderate to high density are obtained, while several other clusters obtained are of very low density. This is attributed to the fact that the proposed algorithm provides highly effective grouping and outlier detection.

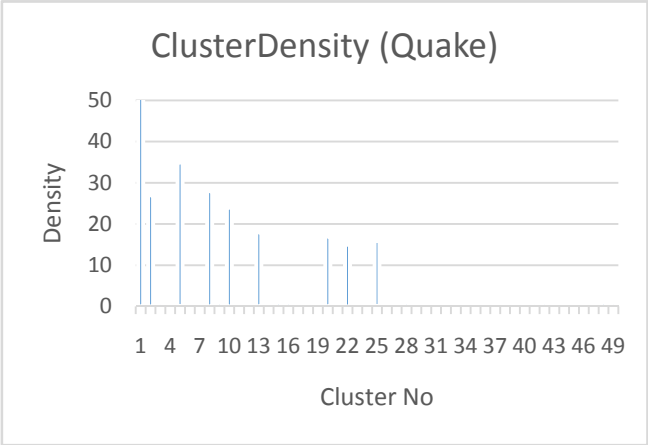

Figure 8. Cluster Density (Quake)

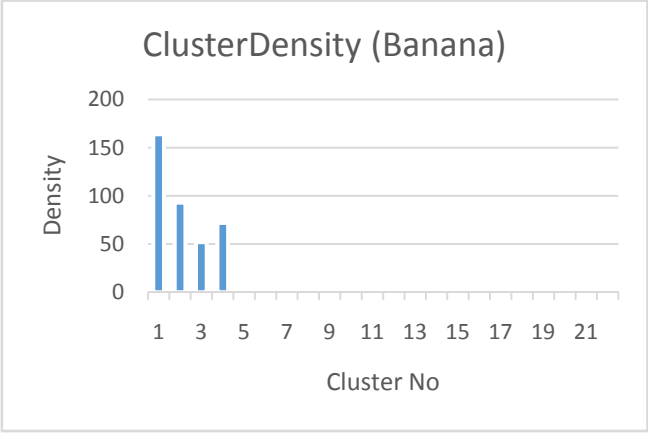

Figure 9. Cluster Density (Banana)

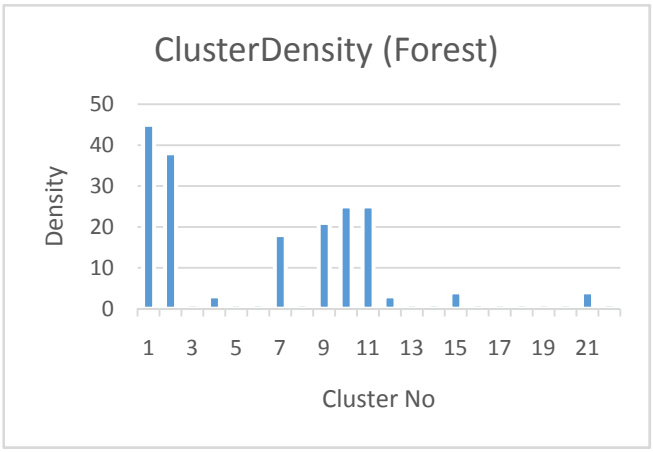

Figure 10. Cluster Density (Forest)

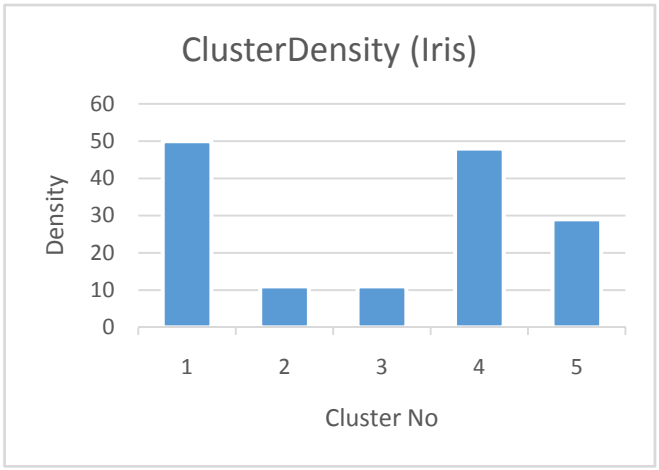

Figure 11. Cluster Density (Iris)

In case of Iris dataset, five clusters were obtained. Among that, two with very high density, one with moderate density and two with low density. It is well known that the Iris dataset has three classes; however, it is a noisy dataset, where two of its classes exhibit high levels of borderline entries. It could be observed that our algorithm is also able to effectively group such data into separate clusters, exhibiting the efficiency of the grouping mechanism in identifying even sub-groupings contained in the datasets.

V. CONCLUSION

This paper presents a density based clustering technique using metaheuristics. The major advantage of the proposed approach is that it can handle huge amounts of data without incurring a high computational cost. Further, density based clustering is often carried out on data with maximum of two dimensions. However, the proposed technique operates on data with much higher dimensions, enabling the technique to be applicable for several varied domains. The results also exhibit the efficiency of the algorithm in identifying outliers; hence this technique can also be used for outlier or noise detection. Future research directions for the proposed technique include parallelization of the selection mechanism to speed up the process and also porting the current code to Big Data environments, hence enabling the algorithm to provide faster results and to handle data of higher dimensions and sizes.

REFERENCES

[1] M. Ester, H.P Kriegel, J. Sander, and X. Xu, "A Density based algorithm for discovering clusters in large spatial databases with noise," Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226-231, 1996.

[2] S.T Mai, X. He, J. Feng, C. Plant, and C. Bohm, "Anytime density based clustering of complex data," Knowledge and Information Systems, vol. 45, no. 2, pp. 319-355, 2015.

[3] A. Azzalini, and G. Menardi, "Density based clustering with non-continuous data," Computational Statistics, vol. 31, no. 2, pp. 771-798, 2016.

[4] S.K. Paul, and P. Bhaumik, "AIDCOR: Artificial immunity inspired density based clustering with outlier removal," International Journal of Machine Learning and Cybernetics, pp. 1-26, 2014.

[5] D. Sengupta, I. Aich, and S. Bandyopadhyay, "Feature selection using feature dissimilarity measure and density based clustering: Application to biological data," Journal of biosciences, vol. 40, no. 4, pp. 721-730, 2015.

[6] Q. Luo, Y. Peng, J. Li, and X. Peng, "MWPCA-ICURD: Density based clustering method discovering specific shape original features," Neural Computing and Applications, pp. 1-12, 2016.

[7] H.P. Kriegel, and P. Martin, "Density based clustering of uncertain data," Proceedings of Knowledge Discovery and Data Mining (KDD), pp. 672-677, 2005.
[8] H.J. Xu, and G.H. Li, "Density based probabilistic clustering of uncertain data," Proceedings of International Conference on Computer Science and Software Engineering, pp. 474-477, 2008.
[9] H. Dirk, B.V. Peter, D. Ralf, and U. Clemens, "Error-aware density based clustering of imprecise measurement values," Proceedings of 7th IEEE International Conference on Data Mining (ICDM), pp. 471-476, 2007.
[10] T. Apinya, and M. Songrit, "U-DBSCAN: A density based clustering algorithm for uncertain objects," Proceedings of 25th International Conference on Data Engineering (ICDE), pp. 136-143, 2010.
[11] J. Sander, M. Ester, H.P. Kriegel, and X. Xu, "Density based clustering in spatial databases: The algorithm GDBSCAN and its applications," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 169-194, 1998.
[12] M. Usman, I. S. Sitanggang, and L. Syaufina, "Hotspot distribution analyses based on peat characteristics using density based spatial clustering," Procedia Environmental Sciences, vol. 24, pp. 132-140, 2015.
[13] K. Nafees Ahmed, T. Abdul Razak, "Fast and effective spatial clustering using multi-start particle swarm optimization technique," International Journal of Engineering and Technology, vol. 8, no. 2, pp. 1229-1237, 2016.