International Journal of Engineering & Technology, 7 (2.9) (2018) 28-32

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Scalable density based spatial clustering with integrated one-class SVM for noise reduction

K. Nafees Ahmed [1] *, T. Abdul Razak [2]

[1] PhD Research Scholar, Dept of CS, Jamal Mohamed College, Tiruchirappalli, Tamil Nadu, India
[2] Associate Professor, Dept of CS, Jamal Mohamed College, Tiruchirappalli, Tamil Nadu, India
**Corresponding author E-mail:nafeesjmc@gmail.com*

Abstract

Information extraction from data is one of the key necessities for data analysis. Unsupervised nature of data leads to complex computational methods for analysis. This paper presents a density based spatial clustering technique integrated with one-class SVM, a machine learning technique for noise reduction, a modified variant of DBSCAN called NRDBSCAN. Analysis of DBSCAN exhibits its major requirement of accurate thresholds, absence of which yields suboptimal results. However, identifying accurate threshold settings is unattainable. Noise is one of the major side-effects of the threshold gap. The proposed work reduces noise by integrating a machine learning classifier into the operation structure of DBSCAN. Further, the proposed technique is parallelized using Spark architecture, thereby increasing its scalability and its ability to handle large amounts of data. Experiments and comparisons with similar techniques indicate high scalability levels and high homogeneity levels in the clustering process.

Keywords: *DBSCAN; One-Class SVM; Noise Reduction; Clustering; Spark.*

1. Introduction

The current Internet age is data rich, but information poor. Meaningful data is the chief requirement of the current age. This has led to an enormous increase in the data processing techniques. However, the major drawback is inherently embedded in the data itself. The distribution of datasets vary and hence a single technique that was designed to process data from a domain would not possibly provide effective results when applied to the data from the same domain, due to variations in the data distribution levels as time progresses. Hence techniques that are as ductile as the data itself are the only ones that can persist.

Unsupervised data processing has always been a challenge due to their unpredictable nature. Ductile and tractable algorithms are the major requirements for processing such data. Domains with such requirements range from image processing to web information processing. The requirements for such processing techniques are constantly on the raise, with the increase in the amount of data being available.

Extracting meaningful information from such data requires grouping them to find commonalities existing between them. This aids in better interpretation of data. Clustering is the process of grouping data such that data within a group/cluster is more cohesive compared to data in different clusters. The process of clustering makes the data meaningful for various applications. Clustering methods are usually classified as partitional and hierarchical clustering techniques. Partitional clustering techniques perform flat clustering based on a single decision criterion such as distance or density. Distance based clustering techniques include K-Means [1] clustering and CLARA [2] to name a few. Density based clustering techniques are currently on the raise due to their flexible operational nature and the stable solution sets generated by them. Density based clustering techniques include DBSCAN [3] and DenClue [4], OPTICS [5] to name a few. Hierarchical clustering algorithms are divided into agglomerative and divisive, corresponding to their basic operational nature, bottom-up or top-down [6].

This paper concentrates on developing a density based spatial clustering technique. DBSCAN, being the precursor of such techniques, is adopted by most of the techniques involving non-uniform clusters. It was identified that DBSCAN has high potency of generating outliers. On further assessment it was identified that several of these outliers effectively fit into the formed clusters. However, they still remain outliers due to the parameters defined for the clustering process. The parameter setting process is always optimal and is identified by the data experts through data analysis and trial and error. As resolving this issue is a complex task, this work presents NRDBSCAN with one-class SVM to reduce the noise levels. Further, a Spark based hybridized DBSCAN has been proposed to improve the scalability levels of the algorithm such that Big Data processing is enabled.

2. Related works

Density based clustering techniques are currently on the increase due to the increase in the requirements for spatial data processing. However, most of these techniques are derivatives from DBSCAN, extending it according to their operating domains. Extensions are usually in terms of reduction in time, parameter automation, input reduction in terms of feature selection and ability to handle varied densities. FDBSCAN [7] and IDBSCAN [8] are two approaches concentrating on the reduction in time component. Both operate by identifying a minimal set of highly representative points for their operation rather than utilizing the entire data points. In addition to this, the IDBSCAN utilized a grid based data selection for reducing the input levels.

Parameters fine-tuning based algorithms include k-VDBSCAN [9], DBCLASD [10] and APSCAN [11]. The k-VDBSCAN is a parameter free technique that automatically identifies the parameters. DBCLASD operates on large data and is a parameter free clustering technique, enabling gradual cluster expansion on the basis of the neighbours and their densities. APSCAN utilizes Affinity Propagation technique to identify the cluster densities based on local data variations.

A density based clustering technique aiding in the discovery of clusters with varying densities within a single dataset was proposed by Zhu et al in [12]. In-general, the input data is considered to contain data distributed in uniform densities. However, some unusual real time data such as population maps tend to contain such data. This leads to a complex issue, as increasing the density levels for the entire process includes several outliers into clusters, while reducing the density levels misses several legitimate clusters. Two approaches exist in literature to solve this issue, namely modifying the algorithm appropriately and rescaling the data. The technique proposed in [12] utilizes the latter by rescaling the data to appropriately identify the clusters. DBSCAN is applied to the rescaled data to identify clusters. A similar technique that identifies varied density based clusters was proposed by Louhichi et al. in [13]. The operational process of this technique is divided into two phases. The first phase identifies the density levels of the input data using the exponential spline technique on the distance matrix. Second phase utilizes the density values determined from the first phase as local threshold parameters to identify clusters.

Scalability issues are least of the explored issues in the density based clustering approaches. Until the recent availability of Big Data, sufficient scalability was provided by the conventional systems itself. However, both computational and storage complexity levels increase when operated in terms of Big Data. The ADBSCAN [14] technique aims to provide scalability to medical and complex data. This technique operates by using a sequence of lower bounding functions to produce several approximate distances. These enable deduction of the actual clusters. Other density based clustering techniques include VDBSCAN [15], GMDBSCAN [16], DDSC [17], EDBSCAN [18] etc.

3. Density based spatial clustering with noise reduction

Density based spatial clustering proposes a grouping mechanism, that operates on the basis of both distance and the node density. The major advantage of using such as approach is that the technique does not rely on centroid based operations, hence the inconsistencies in the formation of clusters are eliminated. Further, density based clustering is an unsupervised approach with no prior information requirements. Density based clustering approaches has the ability to identify clusters of arbitrary shapes and sizes rather than being confined to the traditional circular clusters. It was identified that DBSCAN based techniques exhibits high noise levels. Even though the base clustering process was identified to be efficient, the noise levels generated by DBSCAN were found to be excessive. This paper proposes Noise Reduced DBSCAN (NRDBSCAN), a hybridized density based clustering approach that initially clusters the data, after which it tries to incorporate noise into any of the existing definite clusters, hence reducing the noise levels.

Prior to the actual clustering process, the input data is processed and is converted to the required format. A schema analysis is performed on the data to identify the data types of the contents. The proposed architecture accepts only numerical data, hence textual data are eliminated and categorical and ordinal data are converted to numerical formats. The data pre-processing is followed by the actual clustering process. Algorithm for the proposed NRDBSCAN is shown below.

NRDBSCAN Algorithm

1) Input base data
2) Input threshold levels (minPts, maxDist)
3) Initialize first cluster with a random node n
4) neighbors ← identifyNeighbors(n)
5) If the neighbour count<minPts,
a. Set n into a separate cluster (Outlier)
6) Else expandCluster(n,neighbors)
7) Perform this process until all the nodes are grouped into a cluster
8) For each identified cluster C
a) If density of C < 1
noise ← C
9) Until Noise is not empty
a) For each identified cluster C
i). Prediction ←predict(C,noise)
ii). Add all true predictions to cluster C
iii). Delete corresponding entries from noise
a) if no entries are deleted for the last two itera-tions
i) Allocate new clusters for each entry
i) in noise
ii) Empty noise

function identifyNeighbors (n)
1) For all nodes n1 in data
a) If distance(n,n1)<maxDist
i. Add n1 to neighborList
2) return neighborList

function expandCluster(n,neighbors)
1) Add n to the cluster C
2) For each neighbour n1
b) neighborL1 ← identifyNeighbors(n1)
c) if neighborL1 count >= minPts
i. Add n1 to C

function predict(C,noise)
1) Initialize one-class SVM with polynomial kernel
2) Set the degree of kernel to be the dimensions of C
3) Train SVM with C
4) Predictions ← Apply noise to trained kernel
5) Return Predictions

3.1. Initial level cluster formulation using DBSCAN

The pre-processed data is analyzed and the maximum acceptable distance (maxDist) and the minimum neighbor requirements (minPts) are identified using the data distribution. However, accurately identifying the best parameters is not possible in a single iteration. This is a trial and error based fine-tuning process. Further, these parameters vary with the dataset being used. Hence distribution based transient values are initially identified and the final parameters are identified by methodically increasing and decreasing the parameter values to identify the best parameter set for the current data under analysis.

The process begins with a random node in the dataset. The initial cluster is composed of this single node n. Neighbors of the selected node (n) are identified using maxDist as the threshold. If the node satisfies the minimum neighbor requirements (minPts), it is considered to be a part of a cluster and not an outlier. Hence, this is followed by the cluster expansion phase.

3.2. Cluster expansion

The cluster expansion phase finds the neighbors of each node in n. If each of these nodes satisfies the minPts requirement they are incorporated as an entity in the cluster. This process is continued until all the appropriate nodes are grouped into the cluster. However, several clusters might exist in a dataset. Hence some points would remain outside the composed cluster. A random such point is considered as the base for the next cluster. The neighbor identification and cluster expansion phase are repeated to identify the points corresponding to the new cluster. This process is repeated until all the points are categorized into a cluster.

Though all points are categorized into clusters, not all clusters would be composed with high or even considerable densities. Some clusters would be composed of single nodes. These nodes

are intended to be outliers. Major reasons for the occurrence of outliers are improper parameter tuning, their actual presence or inappropriate node validation in the expansion phase. Though the actual presence of outliers needs to be considered, the other two issues also play a vital role in the occurrence of outliers in DBSCAN. Hence a second level examination would incorporate several nodes that could have been components of the existing clusters, leading to a reduction in the outliers.

3.3. One-class SVM based noise reduction

DBSCAN eliminates several nodes, considering them as outliers. However, they might not necessarily correspond to outliers, rather they could be components of the defined clusters. Utilizing the conventional conditional checks utilized by DBSCAN again ultimately has the same effects. Hence the proposed NRDBSCAN incorporates a machine learning classifier, One-Class Support Vector Machine (SVM) for the categorization process.

Utilizing all the clusters for training a binary/multi-class classifier and then performing classification on the detected outliers has several downsides. The major issue is that binary or multi-class classifiers definitely categorize the data into any one of the groups. Hence all the outliers would definitely be grouped into a cluster, leading to zero outliers. However, the proposed work is intended on grouping only the appropriate nodes and retaining the outliers. Hence a one-class classifier would be the best suited approach. One-class classifiers are a special case of classifiers, where the pattern of a single class would be well known, while the patterns that do not confine to the well-known trained patterns would be considered as outliers.

Output from DBSCAN is in terms of clusters. Data within clusters have obvious associations, hence these can be used as the training data for the classifiers. The noise reduction phase operates by training the one-class SVM using data from one cluster and performing predictions on the detected outliers. Outliers categorized as components of the cluster used for training are incorporated into the cluster. The residual outliers are considered for processing by training the one-class SVM with data from the next cluster. This process is continued for all the defined clusters with considerable density levels.

One-class SVM was suggested by Scholkopf et al. in [19]. It operates by separating the data points from the origin and considering the origin alone as the second class. The base operational nature of one-class SVM is to maximize the distance from the hyper plane created by the data points to the origin. The resultant of this is a binary function that effectively captures the regions in the input space, returning +1 for data contained in the hyper plane and -1 for others. One-class SVM used in the NRDBSCAN utilizes the polynomial kernel [20], as the input data has the tendency to contain several dimensions.

Not all data are expected to be components of the clusters. After the entire process, some residual data remains, which are categorized as outliers.

4. Results and discussion

DBSCAN was implemented in Spark 2.0 and the noise reduction components were incorporated into the operational process. Analysis of the NRDBSCAN was performed by using four benchmark datasets. Clustering specific datasets such as Iris and Banana and spatial datasets such as quake and forest were used for identifying the efficiency of the algorithm.

Cluster densities and the number of clusters obtained from each of the datasets using NRBDSCAN is shown in Figures (1-4). Cluster densities correspond to the number of nodes in each of the formulated cluster. A density of one indicates an outlier. It could be observed that the clusters formulated using the proposed approach exhibits low outlier levels and clusters of considerable densities. Iris and quake, containing low variations exhibit low outlier levels, while forest and banana exhibits high variations, hence high outliers. However, it could be observed that the clusters other than outliers exhibit high density levels.

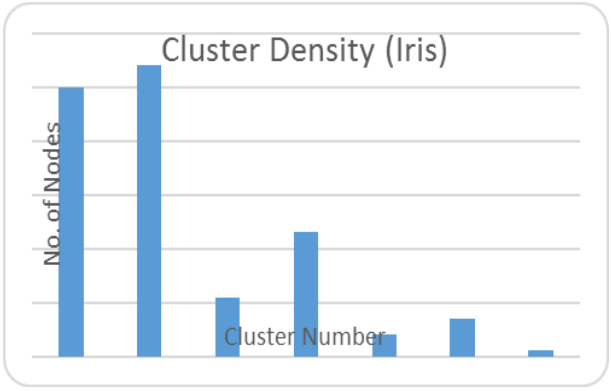

Fig. 1: Cluster Density (Iris).

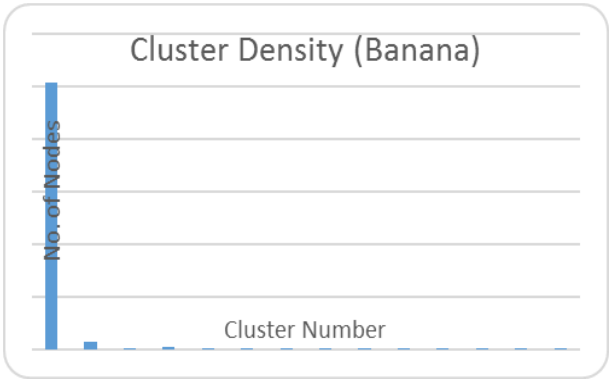

Fig. 2: Cluster Density (Banana).

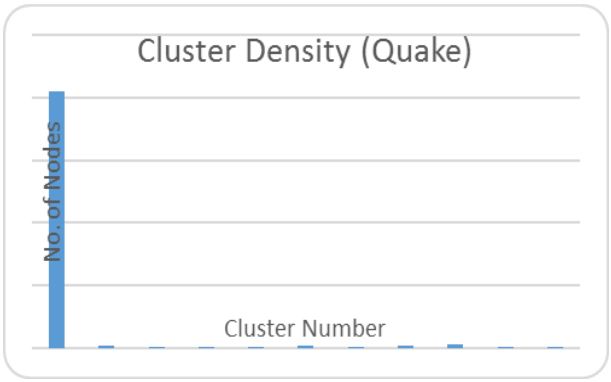

Fig. 3: Cluster Density (Quake).

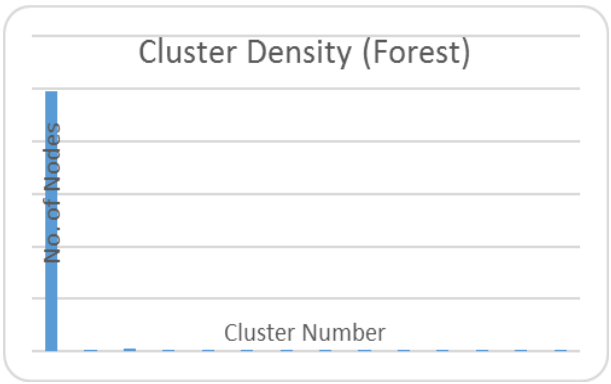

Fig. 4: Cluster Density (Forest).

Intra-cluster radius of the proposed NRDBSCAN is presented in Figures (5-8). It could be observed that the proposed technique exhibits low to moderate intra cluster distance levels. Moderate levels are due to the varied shaped clusters formed by the algorithm.

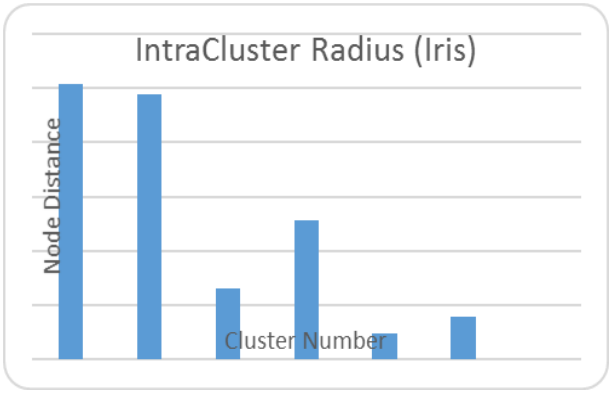

Fig. 5: Intra Cluster Radius (Iris).

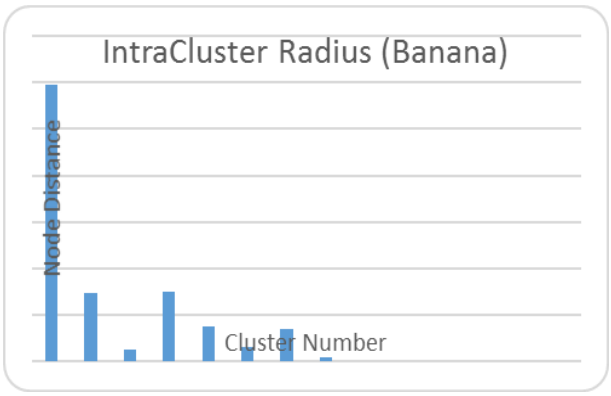

Fig. 6: Intra Cluster Radius (Banana).

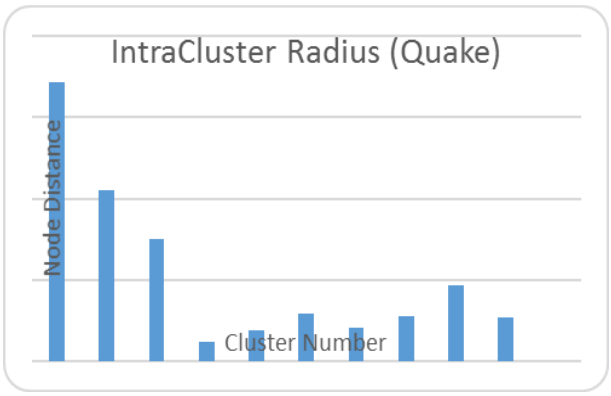

Fig. 7: Intra Cluster Radius (Quake).

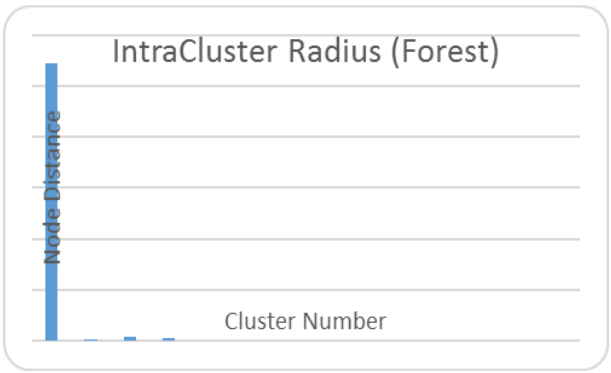

Fig. 8: Intra Cluster Radius (Forest).

Inter-cluster distances exhibited by the algorithm are presented in Figure 9. It could be observed that low inter cluster distances are exhibited by NRDBSCAN. This validates our claim of varied shaped clusters with different density levels.

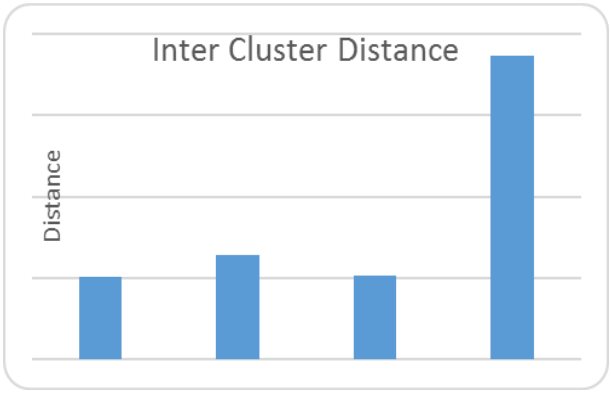

Fig. 9: Inter Cluster Distances.

A comparison in terms of time is carried out between DBSCAN, Modified PSO based DBSCAN [21] and the proposed NRD-BSCAN and is presented in Figure 10. It could be observed that the time taken for modified PSO is the highest. This scalability in terms of time has been achieved by the Big Data and in-memory processing techniques employed using the Spark architecture. However, time taken NRDBSCAN is higher compared to DBSCAN. Even-though that is the case, the difference is in terms of 0.3 sec (maximum). Hence the significance of the time increase is considered to be low.

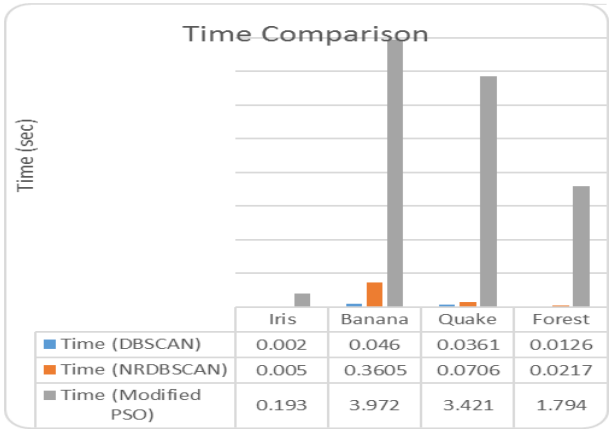

Fig. 10: Time Comparison.

A comparison in terms of the number of clusters generated by DBSCAN and NRDBSCAN is shown in Figure 11. It could be observed that the number of clusters generated by NRDBSCAN is at-least 60% less than the conventional DBSCAN. This exhibits the effective noise reduction levels exhibited by the proposed approach.

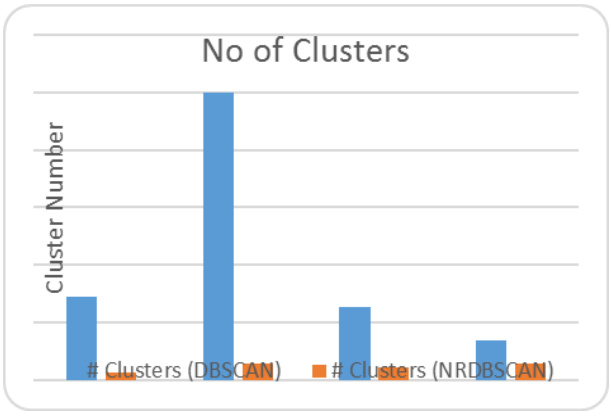

Fig. 11: Cluster Comparison.

5. Conclusion

Clustering, being one of the major techniques for data analysis has seen several adaptations due to the changing constraints. The major adaptations include identifying varied shaped clusters. This paper proposes an enhancement of the DBSCAN that is a pioneer algorithm in identifying varied shaped and varied density clusters. The major downside of DBSCAN was observed to be its requirement for accurate parameter setting. A slightly varied setting results in the increase in outliers. However, identifying the perfect parameter is not feasible. Hence the proposed NRDBSCAN enhances the DBSCAN approach by introducing a noise reduction component based on one-class classifier model. One-class SVM was used to perform this process. Results exhibits significant reduction in the noise levels and effective scalability levels due to the incorporation of Spark based parallelization. Current algorithm operates effectively on fixed density clusters. Future works will concentrate on porting the algorithm to operate on datasets with varied densities and effectively identify clusters with varied density levels.

References

[1] Hartigan JA, and Wong MA,"Algorithm AS 136: A k-means clustering algorithm", *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol.28, No.1, (1979), pp.100-108. https://doi.org/10.2307/2346830.

[2] Wei CP, Lee YH, and Hsu CM,"Empirical comparison of fast clustering algorithms for large data sets", *Proceedings of the 33rd Annual Hawaii International Conference*, (2000), pp:1-10.

[3] Ester M, Kriegel HP, Sander J, and Xu X,"A density-based algorithm for discovering clusters in large spatial databases with noise", *In KDD 1996*, Vol.96, No.34, (1996), pp.226-231.

[4] Hinneburg A, and Keim DA,"An efficient approach to clustering in large multimedia databases with noise", *In KDD 1998*, Vol.98, (1998), pp.58-65.

[5] Ankerst M, Breunig MM, Kriegel HP, and Sander J,"OPTICS: ordering points to identify the clustering structure", *In ACM Sigmod record 1999*, Vol.28, No.2, (1999), pp.49-60. https://doi.org/10.1145/304182.304187.

[6] Güngör E, and Özmen A,"Distance and density based clustering algorithm using Gaussian kernel", *Expert Systems with Applications*, Vol.69, (2017), pp.10-20. https://doi.org/10.1016/j.eswa.2016.10.022.

[7] Zhou S, Zhou A, Jin W, Fan Y, and Qian W,"FDBSCAN: a fast DBSCAN algorithm", *Ruan Jian Xue Bao*, Vol.11, No.6, (2000), pp.735-744.

[8] Tsai CF, and Yeh HF,"Npust: An efficient clustering algorithm using partition space technique for large databases", *Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, (2009), pp: 787-796. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02568-6_80.

[9] Chowdhury AR, Mollah ME, and Rahman MA,"An efficient method for subjectively choosing parameter 'k'automatically in VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise) algorithm", *Proceedings of the 2nd International Conference on Computer and Automation Engineering (ICCAE)*, (2010), Vol.1, pp: 38-41. IEEE.

[10] Parimala M, Lopez D, and Senthilkumar NC,"A survey on density based clustering algorithms for mining large spatial databases", *International Journal of Advanced Science and Technology*, Vol.31, No.1, (2011), pp.59-66.

[11] Chen X, Liu W, Qiu H, and Lai J,"APSCAN: A parameter free algorithm for clustering", *Pattern Recognition Letters*, Vol.32, No.7, (2011), pp.973-986. https://doi.org/10.1016/j.patrec.2011.02.001.

[12] Zhu Y, Ting KM, and Carman MJ,"Density-ratio based clustering for discovering clusters with varying densities", *Pattern Recognition Letters*, Vol.60, (2016), pp.983-997. https://doi.org/10.1016/j.patcog.2016.07.007.

[13] Louhichi S, Gzara M, and Ben-Abdallah H,"Unsupervised varied density based clustering algorithm using spline", *Pattern Recognition Letters*, 2016.

[14] Mai ST, He X, Feng J, Plant C, and Böhm C,"Anytime density-based clustering of complex data", *Knowledge and Information Systems*, Vol.45, No.2, (2015), pp.319-355. https://doi.org/10.1007/s10115-014-0797-0.

[15] Liu P, Zhou D, and Wu N,"VDBSCAN: varied density based spatial clustering of applications with noise", *Proceedings of the International Conference on Service Systems and Service Management*, (2007), pp: 1-4. IEEE. https://doi.org/10.1109/ICSSSM.2007.4280175.

[16] Xiaoyun C, Yufang M, Yan Z, and Ping W,"GMDBSCAN: multi-density DBSCAN cluster based on grid", *Proceedings of the International Conference on e-Business Engineering (ICEBE)*, (2008), pp: 780-783. IEEE. https://doi.org/10.1109/ICEBE.2008.54.

[17] Borah B, and Bhattacharyya DK," DDSC: a density differentiated spatial clustering technique", *Journal of Computers*, Vol.3, No.2, (2008), pp.72-79. https://doi.org/10.4304/jcp.3.2.72-79.

[18] Ram A, Sharma A, Jalal AS, Agrawal A, and Singh R,"An enhanced density based spatial clustering of applications with noise", *Proceedings of the International Conference on Advanced Computing (IACC)*, (2009), pp:1475-1478. IEEE.

[19] Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, and Williamson RC,"Estimating the support of a high-dimensional distribution", *Neural Computation*, Vol.13, No.7, (2001), pp.1443-1471. https://doi.org/10.1162/089976601750264965.

[20] Manevitz LM, and Yousef M,"One-class SVMs for document classification", *Journal of Machine Learning Research*, Vol.2, (2001), pp.139-154.

[21] Nafees Ahmed K, and Abdul Razak T,"Density based clustering using modified PSO based neighbor selection", *International Journal on Computer Science and Engineering (IJCSE)*, Vol.9, No.5, (2017), pp.192-199.