Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 3, June 2019, pp. 1471~1478
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i3.pp1471-1478

OCA: overlapping clustering application unsupervised approach for data analysis

Alvincent E. Danganan[1], Ariel M. Sison[2], Ruji P. Medina[3]
[1, 3]Technological Insttitute of the Philippines, Philippines
[2] Emilio Aguinaldo College, Philippines

Article Info

Article history:

Received Dec 20, 2018
Revised Jan 21, 2019
Accepted Feb 23, 2019

Keywords:

Clustering
K-Means
MAD
Maxdist
Outlier
Overlap

ABSTRACT

In this paper, a new data analysis tool called Overlapping Clustering Application (OCA) was presented. It was developed to identify overlapping clusters and outliers in an unsupervised manner. The main function of OCA is composed of three phases. The first phase is the detection of the abnormal values (outliers) in the datasets using median absolute deviation. The second phase is to segment data objects into cluster using k-means algorithm. Finally, the last phase is the identification of overlapping clusters, it uses maxdis as a predictor of data objects that can belong to multiple clusters. Experimental results revealed that the developed OCA proved its capability in detecting overlapping clusters and outliers accordingly.

Corresponding Author:

Alvincent E. Danganan,
Technological Insttitute of the Philippines,
938 Aurora Blvd., Quezon City, Philippines.
Email: avdanganan836@gmail.com

1. INTRODUCTION

Data mining and knowledge discovery in databases have been an active area of research lately [1]. Data mining applications are useful for commercial and scientific sides [2]. In healthcare application, it is an important method that can be used to detect unknown diseases [3] and identify effective treatments [4]. Data mining technique can be classified into two categories: Supervised and Unsupervised learning [5]. Supervised learning uses datasets that have labels while Unsupervised learning is one of the techniques that can be used to find patterns in unlabeled data sets.

Clustering can be considered unsupervised learning technique. In data mining, clustering is one of the widely use fundamental task [6] and it is used to detect hidden structure or to outine the data category [7]. Clustering aims to find groups from unlabeled data such that all similar data objects is within the same cluster while dissimilar data objects from different cluster [8]. Other study uses overlapping clustering where data objects can belong to multi-cluster.

According to the study [9], most of the real world datasets have overlapping information. Overlapping clustering has been used in many application from wireless sensor network [10] to social network interactions [11]. For example in social network analysis, the overlapping technique is used to detect actors that can belong to multiple communities [12]. Agglomerative hierarchical clustering is another method used to detect overlapping communities in a mobile network [13]. In a wireless sensor network, an energy efficient adaptive overlapping clustering method is established to improve energy efficiency for dynamic continuous monitoring [14]. An algorithm called OverCite which can detect overlapping communities of authors, papers and venues simultaneously using the publication hypergraph and the citation network

information [15]. Another study in network analysis called a density-based link clustering algorithm, its purpose to improve the accuracy of detecting overlapping communities in networks [16].

However, one of many challenging issues are noise or anomalous data, also known as outlier. Having outliers in the dataset may result in inaccurate analysis of data [17], provide a misleading statistical result and may potentially decrease the quality of a data analysis task. Due to this, outlier detection is an important data analysis task, its main objective is to detect anomalous or abnormal data from a given dataset [18].

In this regard, the study is focused on the development of an overlapping clustering application (OCA) that can identify overlapping clusters and outliers respectively. The study considered different research methods and algorithm for the development of the application. One of the algorithm used is the k-means algorithm, because of its simplicity to solve known clustering issues. The study also considered the used of median absolute deviation (MAD), it is known to be one of the most robust measures that are easy to use with the presence of outliers. Maximum distance (maxdist) is another method, it is used to identify data objects assigned to multi-cluster. The OCA application is limited only in handling numerical data.

2. RESEARCH METHOD

2.1. Operational Framework

In this section, the study will demonstrate the workflow of the OCA as shown in Figure 1. In OCA it is necessary that data are converted into a standard spreadsheet and saved it in a csv extension format before loading the data. OCA will then checked the datasets if there are presence of outliers and these identified outliers are removed from the datasets. Then, data objects are clustered accordingly. Afterwards, clusters are checked if there are data objects that overlap within clusters. Finally, the result of the data analysis process is summarized and made available for data interpretation.

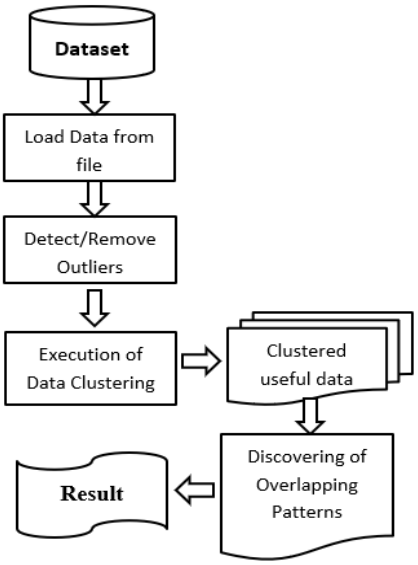

Figure 1. OCA system architecture

2.2. Main Function of OCA

The main function of OCA application consists of three (3) phases.

2.2.1. Phase 1: Outlier Detection using Median Absolute Deviation

Outlier detection aims to find patterns in data that do not conform to expected behavior [19]. Removing and detecting outliers is very important in data mining [20], because it may greatly enhance the performance of statistical technique and data mining algorithms [21]. In order to detect and remove the outliers in the datasets the median absolute deviation (MAD) [22] is used in this study. The process of MAD is discussed in the succeeding section.

To calculate using MAD, all the data objects will be collected and ranked in ascending order. Afterwards, the median value of the series of data objects is to be calculated. Henceforth, the calculated median will be subtracted from each data objects to get the median of absolute deviation. Afterwards, the results are to be sorted in ascending order to determine the median of absolute deviation. Then, the median will be multiplied by b to get the MAD value, where b=1.4826 [23]. In (1) shows the MAD formula.

$$MAD = b\, M_i(\left|x_i - M_j\,(x_j)\right| \quad (1)$$

To determine the outlier, a criterion is computed by median plus or minus threshold value (+/-2, or 2.5, or 3) times the MAD to guide the outlier detection. By default, it is recommended that the threshold value of 2.5 is a reasonable choice for outlier detection. In (2) shows the equivalent criteria value.

$$M + 2.5 * MAD \text{ or } M - 2.5 * MAD \quad (2)$$

All values less than or greater than the computed criterion is considered outliers. This outlier is removed from the datasets before the partition of data objects to form a cluster.

2.2.2. Phase 2: Clustering Using K-Means Algorithm

K-means is one of the oldest and most popular clustering techniques [24]. It is easy to implement and apply even on large data sets [25]. In this section, the researchers discussed how k-means algorithm works.

First, the user enters the number of k clusters, and then the algorithm randomly initializes cluster centroid, one for each cluster. Then, the algorithm calculates the distance of all data objects to the initial centroids using Euclidian distance. Data objects are categorized to its nearest cluster centroid and then cluster centroid is recalculated. This process iterates until the assignments of data objects do not change. In (3) shows the Euclidian distance formula [26].

$$d_{Euclidian}\,(x, y) = \sqrt{\sum_{i=1}^{n}(x_i} - \, y_i)^2 \quad (3)$$

2.2.3. Phase 3: Overlapping Clustering Using Maxdist

In this section, identification of objects to multiple clusters using maxdist [27] is explained. After the formation of clusters using k-means algorithm, calculated distances of each data object assigned on each cluster are saved. The maxdist (maximum distance of an object allowed in a cluster) recorded from each cluster was used as the global threshold in identifying objects that can belong to one or more clusters. Then, the distance of the data object from their respective cluster is calculated to the final centroid of the other cluster. The calculated distance is compared with the maxdist of the other cluster final centroid. If the distance is less than maxdist, then that data object is identified pattern that overlaps with the other final centroid.

In Figure 2, an example of data with three given clusters is shown. To determine whether data object x1 in Cluster 1 overlaps with Cluster 2 the distance of data object x1 is calculated with the final centroid (cent 2) of Cluster 2. Then, the computed distance is compared with data object y3, where y3 is equivalent to the maxdist of Cluster 2. If the distance of x1 is less than the maxdist then x1 is considered data object that overlap with Cluster 2. This method iterates with all the other clusters.

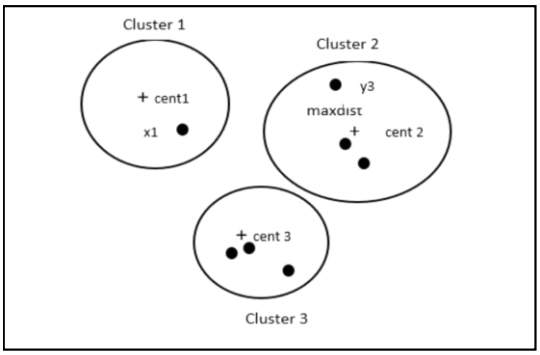

Figure 2. Identification of overlapping patterns

2.3. Visualization Result

The clustering results can be visualized in OCA through a 2-dimensional space or graph. Data objects are characterized by a colored circle dots or points which are represented by a randomly assigned color as a representation for its cluster assignment. While red circle dots or points signifies identified outliers in the datasets. Points that overlap from one cluster to another are circle dots marked with dark border. The cluster centroid (+) is used to represent the composition of clusters. Python programming language was used for the development of the OCA. Figure 3 shows an example of a visualization result window. In Figure 4 shows the result window that displays the detailed information of the data analysis processes.

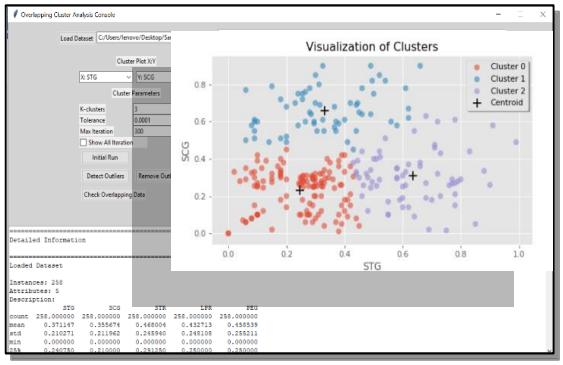

Figure 3. Visualization result

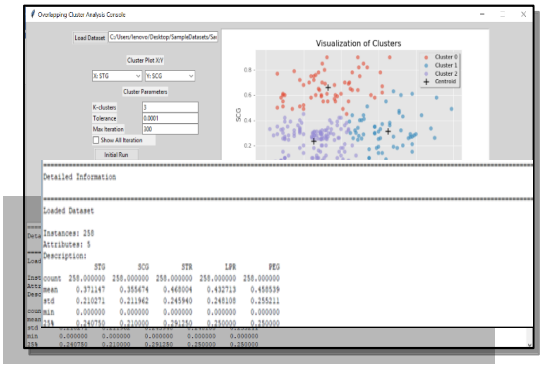

Figure 4. Detailed information result

It provides the number of instances and attributes, the number of data objects assigned on each cluster and the identified outliers from each cluster as well as the overlapping assignment of data.

3. RESULTS AND ANALYSIS

In this section, experiments were conducted to test the developed OCA. The application was implemented using two datasets, synthetic and real datasets.

3.1. Experiment 1

The first experiment used synthetic dataset. The dataset is composed of two numerical attributes with 327 instances. Five were introduced to serve as oultiers data. There are 322 instances that are normal data and 5 instances are outliers.

The data objects are plotted through a 2-dimensional space provided by OCA as shown in Figure 5. First, OCA will used MAD for the identification of outliers in the datasets. Figure 6 shows the visualization result were outliers are identified by OCA. The red circle dots are considered identified outliers in the datasets.

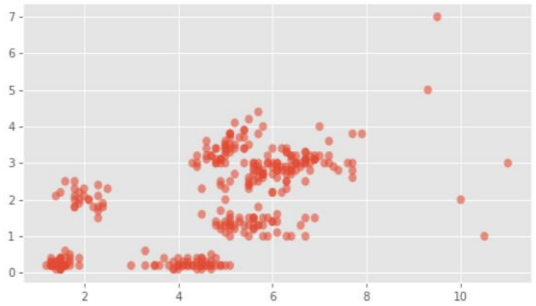

Figure 3. Synthetic datasets scatter plot

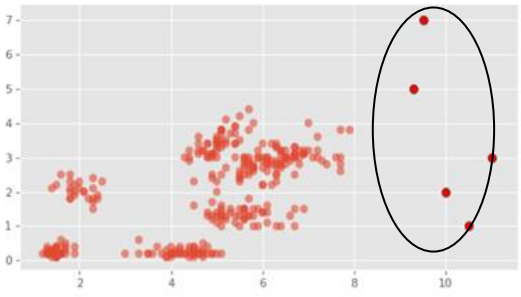

Figure 6. OCA detected outliers

As shown in the visualization result OCA correctly identified all five (5) outliers in the datasets. In Figure 7 shows that outliers are now removed from the datasets.

For the clustering processes, user determines the number of k clusters, wherein the user utilized k=4 in this study. The OCA application takes an input of 4 initial cluster center and each data object is assigned to its nearest cluster center. The clusters in 2-dimensional data space are marked with a randomly assigned color as a representation of the primary belonging of a data objects as shown in Figure 8.

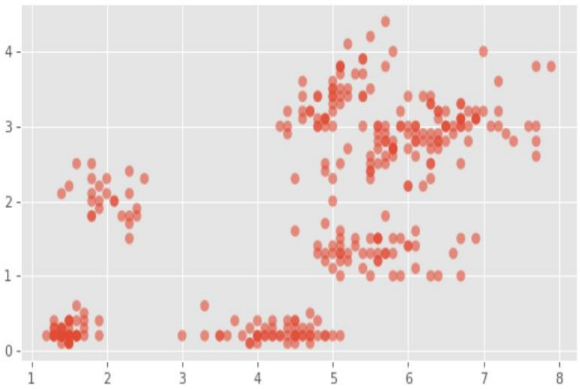

Figure 7. Scatter plot of datasets without outliers

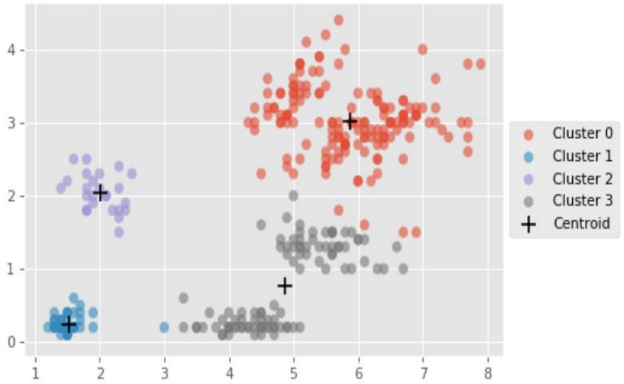

Figure 8. OCA clusters assignment result

Then, OCA will use this maxdist in assigning data objects to multiple clusters. As shown in Figure 9, circle dots marked with dark border are identified data objects that can belong to other clusters. Finally, Tables 1-3 illustrated the detailed information of the experiment done using the synthetic datasets.

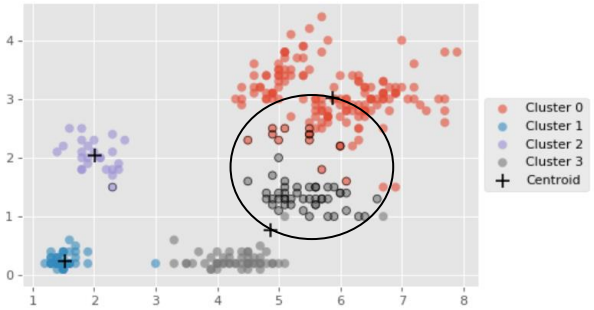

Figure 9. Simulation of data objects that overlap within clusters

Table 1. Implementation Result of Detected Outliers

No. of data Objects	No. of Outliers	Found Outliers
327	5	5

Table 2. Implementation Result of Clustered Data

K=-4	
Clusters	No. of data objects
C0	153
C1	48
C2	25
C3	96

Table 3. Implementation Result of Overlap Clusters

Clusters	Overlaps with C0	Overlaps with C1	Overlaps with C2	Overlaps with C3
C0	-	0	0	13
C1	0	-	0	0
C2	0	1	-	0
C3	45	0	0	-

Experiment 1 results shows that OCA accurately identified outliers in the datasets and was able to discover overlap patterns effectively.

3.2. Experiment 2

In this section, real dataset was obtained from UCI Machine learning repository. The obtained data is the IRIS plants dataset that has 150 instances with 4 (sepal length, sepal width, petal length, petal width) numerical attributes. Figure 10 shows the visualization result under sepal width and sepal length attributes. In Figure 11 shows the outliers in the iris datasets that were identified by the OCA application.

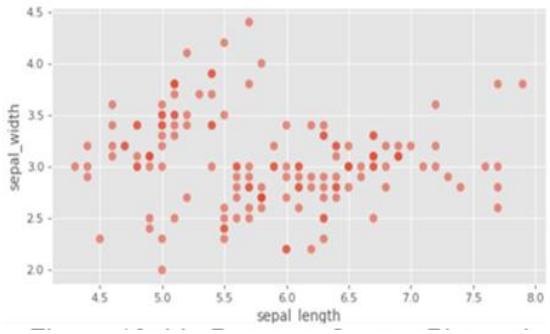

Figure 10. Iris datasets scatter plot under sepal width and sepal length atributes

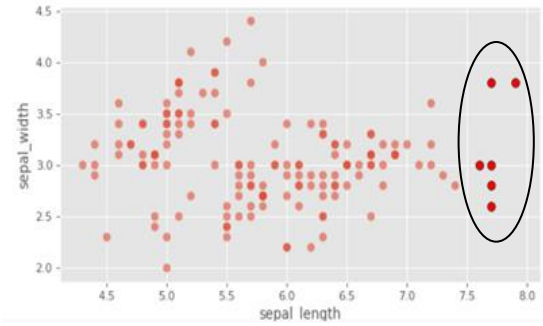

Figure 11. Outliers under sepal width and sepal length

These identified outliers are removed from the datasets. Then, OCA takes an input of k from the user, where k=3 clusters as shown in Figure 12. Finally, the identification of overlap patterns as shown in Figure 13.

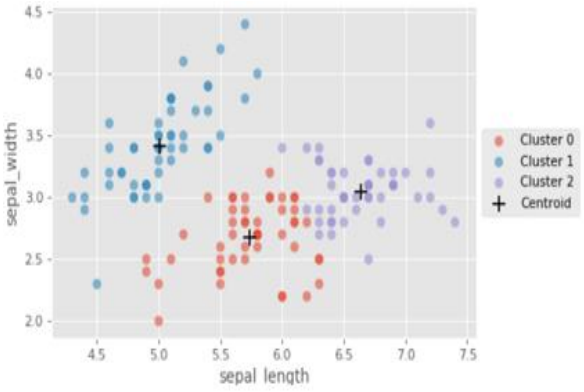

Figure 12. OCA clusters assignment result under sepal width and sepal length

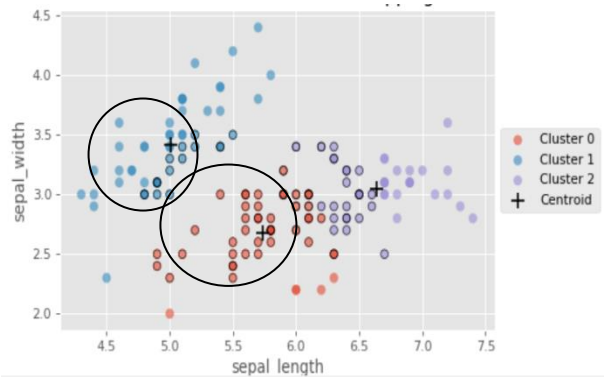

Figure 13. OCA clusters assignment result under sepal width and sepal length

Another experiment where conducted, at this stage the used of petal width and petal length attributes of the iris dataset. The following simulated results are shown in Figure 14.

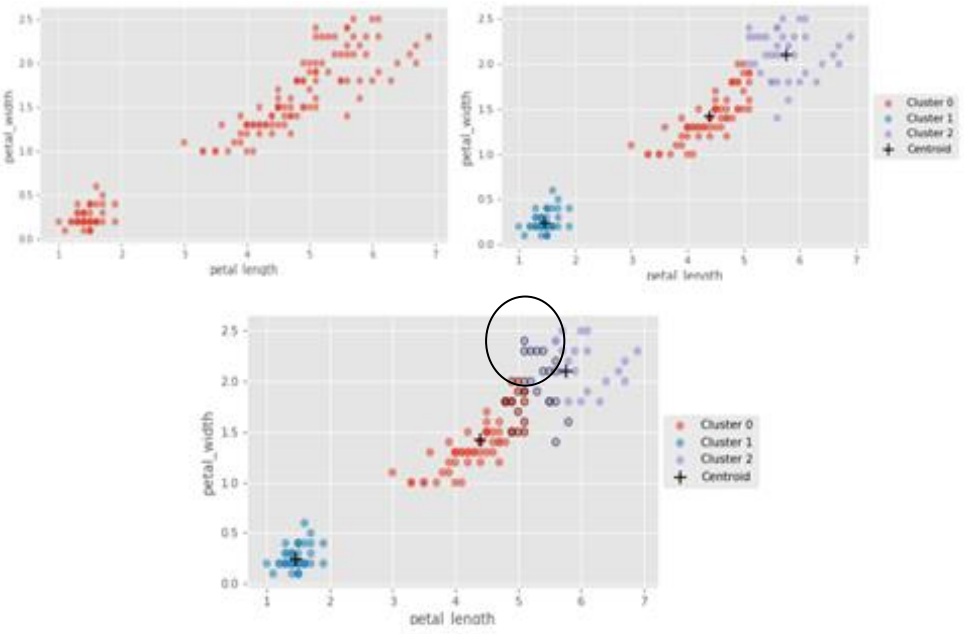

Figure 14. Simulation result under petal width and petal length attributes

In Table 4 illustrated the detailed information of the experiment done using the four (4) numerical attributes (sepal width, sepal length, petal width, petal length) of the iris plant datasets. The experimental result shows that OCA found a total of 6 outliers out of 150 instances under sepal width and sepal length while none in petal width and petal length. For the identification of overlapping patterns under sepal width and sepal length, a total of 77 identified patterns out of 150 instances that overlapped between clusters while in petal width and petal length a total of 34 patterns were identified.

Table 4. Implementation Result Under Iris Plant Dataset

Attributes	Outliers	Clusters	No. of Data Objects	Overlap with C0	Overlap with C1	Overlap with C2
SEPAL		C0	49	-	38	15
Width and	6	C1	50	0	-	0
Length		C2	45	24	0	-
PETAL		C0	63	-	0	17
Width and	0	C1	50	0	-	0
Length		C2	37	17	0	-

Based from the above results, the developed OCA prove its capability to provide better identification of clusters that overlap and outliers accordingly.

4. CONCLUSION AND FUTURE WORKS

The study presented an overlapping clustering application or OCA for data analysis. Based on the experimental results, the developed OCA demonstrated its capability in terms of detecting the abnormal values (outliers) and identification of clusters with overlaps. OCA is very useful data analysis tool for outlier detection analysis, data clustering and detection of overlapping clusters.

Despite providing a good result, it is recommended that more tests need to be done. The developed OCA only works with numerical datasets; therefore, modification of the application can be considered for future works. Furthermore, it is recommended that an alternative approach, which is not sensitive to the random initialization of cluster center, be considered as future study.

REFERENCES

[1] A. R. S. Ritu, M. Afshar Alam, "K-Means Clustering in Spatial Data Mining using Weka Interface," *Int. J. Comput. Appl.*, pp. 13–16, 2015.

[2] P. G. Subbarao, P. S. Khan, and K. V. Kumar, "Case Study on Data Mining Application in Health Care Monitoring Systems," *Res. Inven. Int. J. Eng. Sci.*, vol. 6, no. 5, pp. 79–82, 2016.

[3] R. A. Haraty, M. Dimishkieh, and M. Masud, "An Enhanced *k* -Means Clustering Algorithm for Pattern Discovery in Healthcare Data," *Int. J. Distrib. Sens. Networks*, vol. 11, no. 6, p. 615740, 2015.

[4] P. Kalyani, "Approaches to Partition Medical Data using Clustering Algorithms," *Int. J. Comput. Appl.*, vol. 49, no. 23, pp. 975–8887, 2012.

[5] S. Garg and A. Sharma, "Comparative Analysis of Data Mining Techniques on Educational Dataset," *... J. Comput. Appl. ...*, vol. 74, no. 5, pp. 2–6, 2013.

[6] Y. Hou, J. J. Whang, D. F. Gleich, W. Lafayette, and W. Lafayette, "Non-exhaustive, Overlapping Clustering via Low-Rank Semidefinite Programming Categories and Subject Descriptors," *ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, no. Section 3, pp. 427–426, 2015.

[7] M. Alaqtash, M. A. Fadhil, and A. F. Al-azzawi, "A Modified Overlapping Partitioning Clustering Algorithm for Categorical Data Clustering," *Bull. Electr. Eng. Informatics*, vol. 7, no. 1, 2018.

[8] A. Rezgui, C. N. Cir, and N. Essoussi, "Overlapping Clustering with Outliers Detection," in *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods*, 2014, pp. 279–286.

[9] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," *Expert Syst. Appl.*, vol. 67, pp. 12–18, 2017.

[10] C. Ma, L. Wang, J. Xu, Z. Qin, L. Shu, and D. Wu, "An Overlapping Clustering Approach for Routing in Wireless Sensor Networks," in *2013 IEEE Wireless Communications and Networking Conference (WCNC): SERVICES & APPLICATIONS An*, 2013, pp. 4375–4380.

[11] S. Y. Bhat and M. Abulaish, "A density-based approach for mining overlapping communities from social network interactions," *Proc. 2nd Int. Conf. Web Intell. Min. Semant. - WIMS '12*, p. 1, 2012.

[12] M. E. Celebi, *Overview of overlapping Partitional clustering methods*, no. January. 2015.

[13] P. Kim and S. Kim, "A detection of overlapping community in mobile social network," *Proc. 29th Annu. ACM Symp. Appl. Comput. - SAC '14*, pp. 175–179, 2014.

[14] Y. Hu, Y. Niu, J. Lam, and Z. Shu, "An Energy-Efficient Adaptive Overlapping Clustering Method for Dynamic Continuous Monitoring in WSNs," *IEEE Sens. J.*, vol. 17, no. 3, pp. 834–847, 2017.

[15] T. Chakraborty and A. Chakraborty, "OverCite: Finding Overlapping Communities in Citation Network," in *International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 2013.

[16] X. Zhou, Y. Liu, J. Wang, and C. Li, "A density based link clustering algorithm for overlapping community detection in networks," *Phys. A Stat. Mech. its Appl.*, vol. 486, pp. 65–78, 2017.

[17] A. E. Danganan, A. M. Sison, and R. P. Medina, "An Improved Overlapping Clustering Algorithm to Detect Outlier," *Indones. J. Electr. Eng. Informatics*, vol. 6, no. 4, pp. 401–409, 2018.

[18] M. Ahmed and A. N. Mahmood, "A novel approach for outlier detection and clustering improvement," *Proc. 2013 IEEE 8th Conf. Ind. Electron. Appl. ICIEA 2013*, pp. 577–582, 2013.

[19] K. Singh and S. Upadhyaya, "Outlier Detection: Applications And Techniques.," *Int. J. Comput. ...*, vol. 9, no. 1, pp. 307–323, 2012.

[20] M. Mansur, M. Sap, and M. Noor, "Outlier Detection Technique in Data Mining: A Research Perspective," *Inf. Syst.*, pp. 23–31, 2005.

[21] J. J. Manoharan, "Outlier Detection Using Enhanced K-means Clustering Algorithm and Weight Based Center Approach," *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 4, pp. 453–464, 2016.

[22] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013.

[23] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *J. Am. Stat. Assoc.*, vol. 88, no. 424, pp. 1273–1283, 1993.

[24] U. R. Raval and C. Jani, "Implementing & Improvisation of K-means Clustering Algorithm," *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 5, pp. 191–203, 2016.

[25] P. Thi, T. Binh, T. N. Le, and N. P. Xuan, "Advanced SOM & K Mean Method for Load Curve Clustering," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, pp. 4829–4835, 2018.

[26] P. A. Zizwan, M. Zarlis, E. B. Nababan, I. Singh, and P. Dwivedi, "K-Means Algorithm Performance Analysis With Determining The Value Of Starting Centroid With Random And KD-Tree Method," in *Journal of Physics*, 2017.

[27] S. Baadel, F. Thabtah, and J. Lu, "MCOKE : Multi-Cluster Overlapping K-Means Extension Algorithm," *Int. J. Comput. Electr. Autom. Control Inf. Eng.*, vol. 9, no. 2, pp. 427–430, 2015.