

Overlapping clustering with k-median extension algorithm: An effective approach for overlapping clustering

Alvincent E. Danganan¹, Regina P. Arceo²

¹College of Computer Studies, Tarlac State University, Tarlac, Philippines

²Department of Information Technology, College of Computer Studies, Tarlac State University, Tarlac, Philippines

Article Info

Article history:

Received Dec 19, 2021

Revised Mar 22, 2022

Accepted Apr 5, 2022

Keywords:

Clustering

Kmeans

Kmedian

Outlier

Overlapping

ABSTRACT

Most natural world data involves overlapping communities where an object may belong to one or more clusters, referred to as overlapping clustering. However, it is worth mentioning that these algorithms have a significant drawback. Since some of the algorithm uses k-means, it also inherits the characteristics of being noise sensitive due to the arithmetic mean value which noisy data can considerably influence and affects the clustering algorithm by biasing the structure of obtained clusters. This paper proposed a new overlapping clustering algorithm named OCKMEx, which uses k-median to identify overlapping clusters in the presence of outliers. This new method aims to determine the insensitivity of the OCKMEx algorithm in locating data points that overlap even with outliers. An experimental evaluation of the algorithm was conducted wherein synthetic datasets served as a data source, and the F1 measure criterion was applied to assess the OCKMEx algorithm performance. Results indicate that the OCKMEx algorithm implementing the use of k-median performed a higher accuracy rate of 100% in identifying data points that overlap even with outliers compared to the existing k-means algorithm. The algorithm exhibited a promising performance in identifying overlapping clusters and was resistant to outliers.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Alvincent E. Danganan

College of Computer Studies, Tarlac State University

Tarlac, Philippines

Email: avdanganan@tsu.edu.ph

1. INTRODUCTION

With the rise of Information technology also comes the upward movement of data in several fields [1]. Information technology and database research have served as a bridge to develop an approach to manipulate and store data which plays a significant role in coming up with important decisions in an organization or business. The procedure in which hidden knowledge is extracted from volumes of raw data using algorithms and techniques from statistics, machine learning, and database management system is referred to as data mining. Data mining in extensive data enables organizations and firm decisions by assembling, accumulating, and accessing corporate data [2]. In healthcare, a variety of of algorithms are developed and used [3] that cultivate the existing information into a useful data.

Data mining is gaining popularity in disparate research fields due to its boundless applications and approaches to mine the data in an appropriate manner [4]. One such data mining technique is Clustering. Clustering refers to a set of objects grouped so that the things in the same group are more similar in some particular manner to each other compared with those in the other group [5]. Multiple research areas apply this technique, specifically data mining, statistical data analysis, machine learning, pattern recognition, image

analysis, and information retrieval [6]. Clustering serves as a significant area in data mining applications and data analysis. It is a specific operation that must be a proper subset of the other [7].

Most of the clustering algorithms generate exclusive clusters where each item could be a part of a single cluster only. In the field of medical datasets, specifically real-world data, this contains inherently overlapping information, which applies the method of overlapping clustering that permits one item to be a part of more than one cluster [5]. Another essential data mining issue is outlier detection, which identifies and removes data objects from a given data set. Outlier detection remains to be a research branch in data mining which plays a significant and extensive role because of its widespread use in a wide range of applications [8]. The outlier is the data item whose value falls outside the bounds in the sample data may indicate anomalous data [6]. Moreover, it is a data item whose values vary from the rest of the data or whose values fall outside the described range [9]. The detection of outliers translates to information that is significant and actionable in a wide variety of applications such as fraud detection [10], [11], intrusion detection in cybersecurity [12], and health diagnosis [13]. Like in the medical world, a normal body temperature fluctuation might signify an outlier [14]. Finding anomalous points among the data points is the basic idea to find out an outlier. Outlier detection is a significant research problem that intends to locate data objects that are considerably different, exceptional, and inconsistent in the database [15].

With this concern, the researchers introduced a new overlapping clustering with a k-median extension algorithm (OCKMEx). In statistics, the median is incredibly resistant to outliers. To deter the median from the bulk of the information requires at least 50% of the data to be contaminated [16]. Through its use as the primary factor in the placement of cluster centers, k-medians can assimilate the robustness that the median provides. This new method aims to determine the insensitivity of the OCKMEx algorithm in locating data objects that overlaps in a cluster with the influence of outliers.

2. METHOD

2.1. Overlapping clustering with k-median extension algorithm

This section explains the algorithm OCKMEx by detecting data points that overlap within clusters. OCKMEx is a new overlapping clustering algorithm that discovers the assignment of data points into more than one cluster. Maximum distance (maxdist) is applied, which acts as a global threshold in assigning objects to multiple groups. OCKMEx comprises two separate phases, as shown in Figure 1.

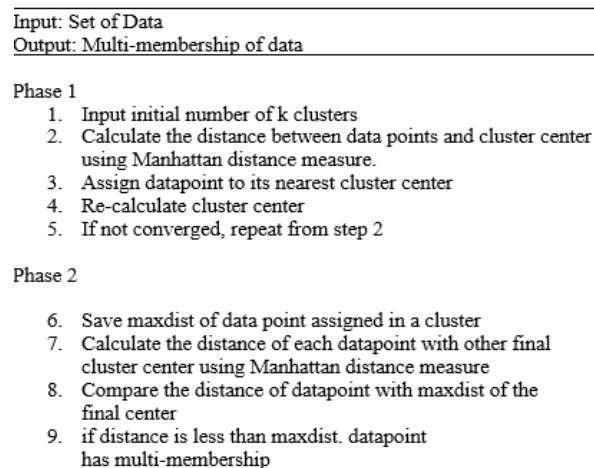


Figure 1. OCKMEx algorithm

PHASE 1: Segmentation of data points to form a cluster is done using the k-median algorithm. The objective of the k-median clustering algorithm is to find the distance between data points to its nearest cluster center using the 1-norm distance, which is called the Manhattan measure. First, the algorithm accepts the initial input of the k center as the initial representation of k points or the k median. Then, the algorithm assigns every point to its nearest median. The algorithm re-calculates the median using the median of each feature. This process iterates until the convergence criterion achieves the desired properties. The criterion function for the k-median algorithm [17] is defined in (1).

$$D(x, y) = \sum |x_i - y_i| \tag{1}$$

PHASE 2: Maxdist [18] identifies each data object and the cluster centers' distance. Maxdist is saved after the first actual run of k-median, which will also serve as a global threshold in determining the multiple memberships of data points in a cluster. A membership table was generated, including data object vectors designated to each cluster and their final cluster centroid. A value of 1 or 0 will be assigned to each data object in the membership table, denoting that it is a member or not a member of a cluster. The final k center's distance from the other cluster and the data points within another group is computed through iteration. Then maxdist of a cluster will now be compared with the calculated distance. Suppose the result of the distance comes up with a lesser value of the maxdist. In that case, the data object is assigned as a member of that cluster centroid, updating the membership table with a value of 1 representing cluster membership.

2.2. Accuracy

The researchers used Recall, Precision, and F-measure to evaluate the accuracy of overlapping clustering results. Precision is the fraction of correctly identified pairs in the same cluster, and recall is the fraction of actual pairs that were identified. The formula for precision and recall is shown in [19].

$$Precision = \frac{Number\ of\ Correctly\ Identified\ Linked\ Pairs}{Number\ of\ Identified\ Linked\ Pairs} \tag{2}$$

$$Recall = \frac{Number\ of\ Correctly\ Identified\ Linked\ of\ Pairs}{Number\ of\ True\ Linked\ Pairs} \tag{3}$$

To model the desired precision and recall, the F-measure, also referred to as the F1 score combining precision and recall, was also used. F-measurement calculates the weighted harmonic mean of recall and precision [20]. The higher the value for the F-measure, the better the detection accuracy, where 0 represents the worst and 1 illustrates a perfect detection [21]. The formula below defines the calculation of F-measure.

$$F_1\ Score = \frac{2 * RECALL * PRECISION}{PRECISION + RECALL} \tag{4}$$

3. RESULTS AND DISCUSSION

In this section, two experiments were conducted to test the developed OCKMex algorithm. The first experiment identifies if OCKMex algorithm can group data into clusters and locate points that belongs to multiple clusters. On the other hand, the second experiment aims to examine OCKMex algorithm's accuracy performance in locating data points that overlap within clusters with outliers in comparison with the existing algorithm MCOKE. Synthetic datasets were used for the implementation of the algorithm.

3.1. Experimentation 1

The aim of experiment 1 is to test whether the OCKMEx algorithm can process the grouping of data samples into a cluster based on its similarity features and locate data points that belong to multiple memberships. The study used synthetic datasets with two attributes (Rating, Absences) with 20 instances with two considered linked pairs. Table 1 shows the experimental synthetic datasets.

Table 1. Synthetic datasets

Student	Rating	Absences	Student	Rating	Absences
Student 1	80	2	Student 11	72	7
Student 2	90	2	Student 12	71	6
Student 3	77	3	Student 13	82	2
Student 4	70	5	Student 14	83	2
Student 5	75	3	Student 15	95	1
Student 6	72	6	Student 16	90	1
Student 7	73	7	Student 17	75	6
Student 8	80	3	Student 18	70	8
Student 9	90	2	Student 19	84	2
Student 10	79	4	Student 20	83	3

3.1.1. Phase 1

In the study's first phase, the synthetic dataset was run with the k-median algorithm to segment the datasets into clusters. A user enters the number of K points as the initial cluster center assigned to each data

point. In this experiment, OCKMex takes an input of 2 clusters center and sets a data point to its nearest cluster center using Manhattan distance measure. Figure 2 shows the simulation result of two clusters in 2-dimensional spaces.

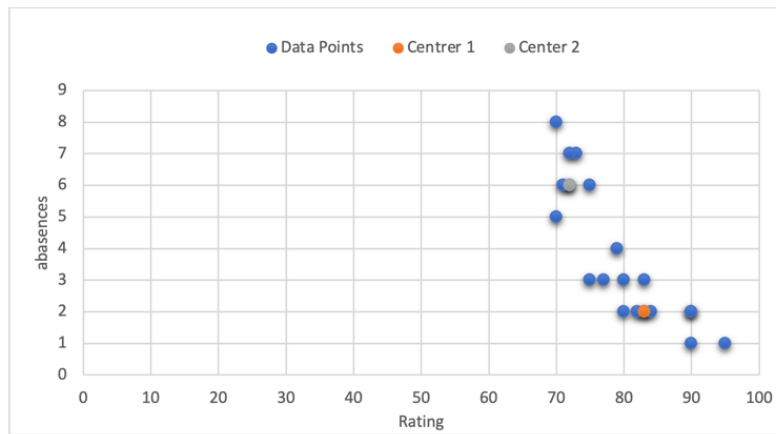


Figure 2. Simulation results of two clusters

After the initial run of OCKMex, an membership table (MT) is generated, showing the vectors of all assigned data points and their cluster center. Each data point in MT represents a value of 1 or 0, which signifies its membership in the cluster. If a data point is assigned one, it means membership to that cluster; otherwise, 0 for non-membership. Table 2 shows the result of the MT obtained from the produced cluster using the k-median algorithm.

Table 2. Two clusters membership table

Vectors	Cluster Center	Cluster 1	Cluster 2
0, 2	83.0,2.0	1	0
90, 2	83.0,2.0	1	0
77, 3	83.0,2.0	1	0
70, 5	72.0,6.0	0	1
75, 3	72.0,6.0	0	1
72, 6	72.0,6.0	0	1
73, 7	72.0,6.0	0	1
80, 3	83.0,2.0	1	0
90, 2	83.0,2.0	1	0
79, 4	83.0,2.0	1	0
72, 7	72.0,6.0	0	1
71, 6	72.0,6.0	0	1
82, 2	83.0,2.0	1	0
83, 2	83.0,2.0	1	0
95, 1	83.0,2.0	1	0
90, 1	83.0,2.0	1	0
75, 6	72.0,6.0	0	1
70, 8	72.0,6.0	0	1
84, 2	83.0,2.0	1	0
83, 3	83.0,2.0	1	0
Total Count		12	8

3.1.2. Phase 2

In this phase, the researchers used each cluster's maximum distance (maxdist) as a threshold in discovering the belonging of data points to multiple groups. This time, the OCKMex algorithm iterates to calculate the distance of data points assigned to its primary cluster with another cluster center. If the estimated distance of the data points is less than maxdist, MT is modified to 1; otherwise, 0. As shown in Table 3, two (2) instances overlap with another cluster.

3.2. Experimentation 2

This experiment aims to examine the accuracy performance of the OCKMEx algorithm in locating data points that overlap within clusters with outliers compared with the existing algorithm MCOKE [22]. The same synthetic data samples were used based on the first experiment to test the performance of the two algorithms. It comprises two attributes (Rating, Absences), with 25 instances. Five outliers are intentionally integrated into the data samples; thus, 20 instances are normal data with two linked pairs, and five are considered an anomaly in the datasets (Student 21 to Student 25). Table 4 shows the data samples.

Table 3. OCKMEx overlapping results

Vectors	Cluster Center	Cluster 1	Cluster 2
80, 2	83.0,2.0	0	0
90, 2	83.0,2.0	0	0
77, 3	83.0,2.0	0	0
70, 5	72.0,6.0	0	0
75, 3	72.0,6.0	0	1
72, 6	72.0,6.0	0	0
73, 7	72.0,6.0	0	0
80, 3	83.0,2.0	0	0
90, 2	83.0,2.0	0	0
79, 4	83.0,2.0	0	0
72, 7	72.0,6.0	0	0
71, 6	72.0,6.0	0	0
82, 2	83.0,2.0	0	0
83, 2	83.0,2.0	0	0
95, 1	83.0,2.0	0	0
90, 1	83.0,2.0	0	0
75, 6	72.0,6.0	0	1
70, 8	72.0,6.0	0	0
84, 2	83.0,2.0	0	0
83, 3	83.0,2.0	0	0
Total Overlap Count		0	2

Table 4. Synthetic dataset with outliers

Student	Rating	Absences
Student 1	80	2
Student 2	90	2
Student 3	77	3
Student 4	70	5
Student 5	75	3
Student 6	72	6
Student 7	73	7
Student 8	80	3
Student 9	90	2
Student 10	79	4
Student 11	72	7
Student 12	71	6
Student 13	82	2
Student 14	83	2
Student 15	95	1
Student 16	90	1
Student 17	75	6
Student 18	70	8
Student 19	84	2
Student 20	83	3
Student 21	138	9
Student 22	135	7
Student 23	140	8
Student 24	125	4
Student 25	127	6

In this experiment, two tests were conducted wherein one approach included the OCKMEx algorithm, and the other was the MCOKE algorithm. These two algorithms used different techniques in arranging data points to form a cluster. OCKMEx algorithm uses k-median (minimization of 1-norm distance). In contrast, k-means (minimization of 2-norm distance) [16] was applied for the MCOKE algorithm, but both algorithms use maxdist to detect data points that overlap within clusters. For the first experimental test, the OCKMEx algorithm inputs two clusters k center for the initial formation of clusters. Figure 3 shows the simulated data samples with outliers that were plotted through 2-dimensional space. As seen in the simulated results from Iteration 1 to Iteration 6 in Figure 4, the implementation of k-median in the algorithm is highly immune to outliers' influence to dissuade the k-centers away from the standard data samples.

Iteration 1 in Figure 4(a) shows that the application of k-median algorithm brings high immunity to outliers allowing to obtain members of each cluster. In Figure 4(b) shown minimal movement on cluster center is seen in iteration 2 where the use of the algorithm provides resistance to outliers allowing the formation of clusters. Identification of cluster is still evident in Iteration 3 even in the presence of outliers in Figure 4(c). As shown in Figure 4(d) iteration 4 depicts the insensitivity of the algorithm in locating data objects with the influence of outliers. Iteration 5 demonstrates a minor change with the movement of the cluster center allowing the assignment of data points to clusters in Figure 4(e) and Figure 4(f) shown iteration 6 shows no movement on cluster center displaying consistent resistance to outliers.

The next step is to determine the multi membership of the data points. Tables 5 and 6 shows the membership table as the first test results conducted using the OCKMEx algorithm. The test indicates that OCKMEx eventually detected two (2) instances in the data samples as members of two clusters. The same data samples were processed; this time MCOKE was used. Figure 5 shows how highly vulnerable is MCOKE with the existence of outliers. Based on the simulation results starting from Iteration 1 to Iteration 3, outliers can drastically pull one of the k centers values away from the rest of the expected data, making all outliers a cluster of outliers.

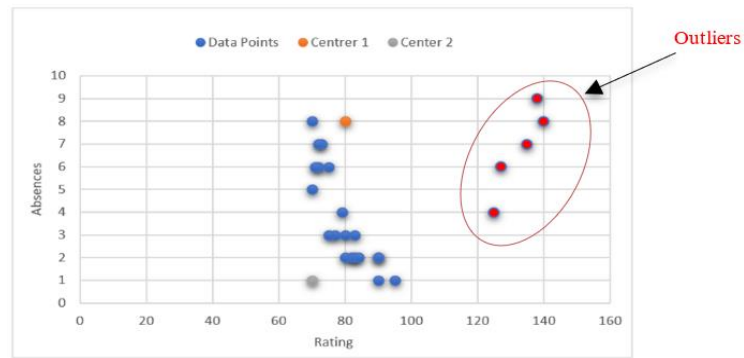


Figure 3. Initial formation of clusters with outliers

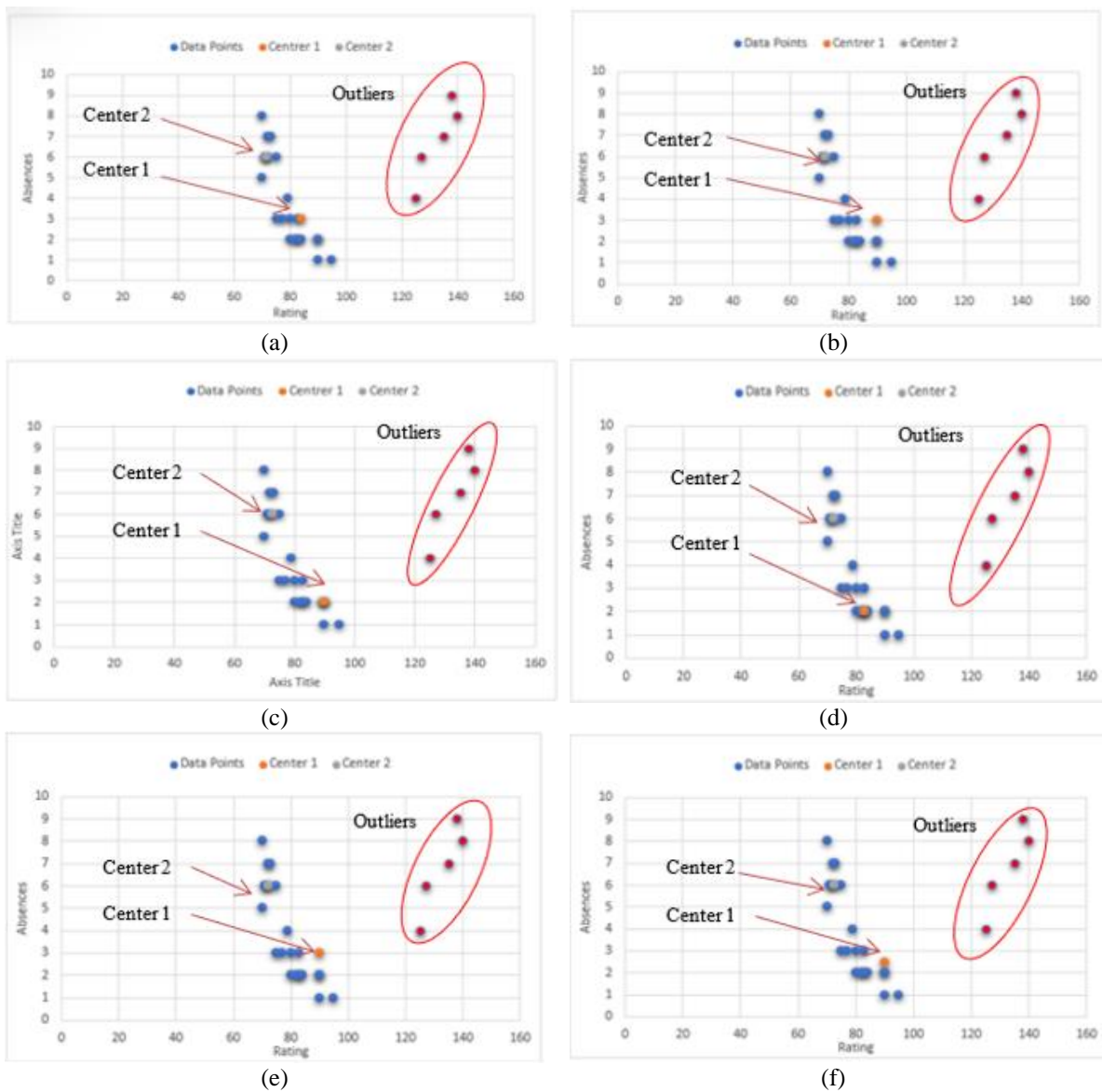


Figure 1. OCKMEx simulation results with outliers for: (a) Iteration 1, (b) Iteration 2, (c) Iteration 3, (d) Iteration 4, (e) Iteration 5, and (f) Iteration 6

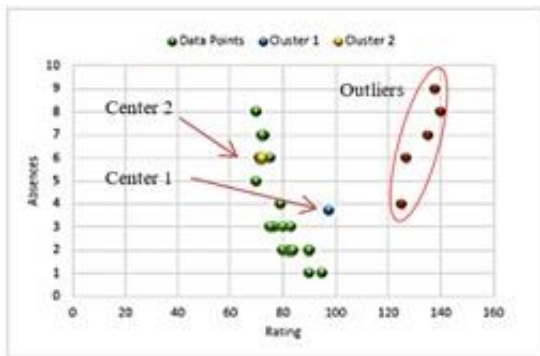
Iteration 1 in Figure 5(a) shows a visible movement of cluster center to the normal data points. In Figure 5(b) shown cluster center is drastically drawn from normal data points displaying sensitivity to outliers in Iteration 2. Iteration 3 shows how highly vulnerable MCOKE is with the presence of outliers as new cluster of outliers was formed with the cluster center evidently pulled from the normal data as shown in Figure 5(c).

Table 5. OCKMEx MT with outliers

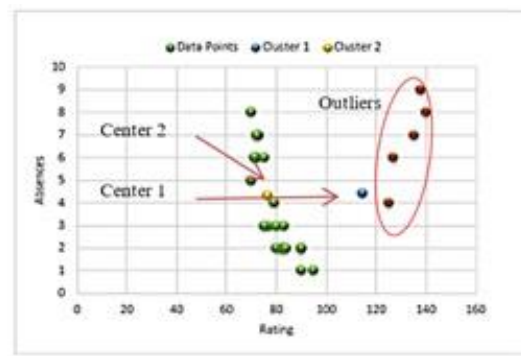
Vectors	Cluster Center	Cluster 1	Cluster 2
80, 2	90.0,2.0	1	0
90, 2	90.0,2.0	1	0
77, 3	73.0,6.0	0	1
70, 5	73.0,6.0	0	1
75, 3	73.0,6.0	0	1
72, 6	73.0,6.0	0	1
73, 7	73.0,6.0	0	1
80, 3	73.0,6.0	0	1
90, 2	90.0,2.0	1	0
79, 4	73.0,6.0	0	1
72, 7	73.0,6.0	0	1
71, 6	73.0,6.0	0	1
82, 2	90.0,2.0	1	0
83, 2	90.0,2.0	1	0
95, 1	90.0,2.0	1	0
90, 1	90.0,2.0	1	0
75, 6	73.0,6.0	0	1
70, 8	73.0,6.0	0	1
84, 2	73.0,6.0	0	1
83, 3	90.0,2.0	1	0
138, 9	90.0,2.0	1	0
135, 7	90.0,2.0	1	0
140, 8	90.0,2.0	1	0
125, 4	90.0,2.0	1	0
127, 8	90.0,2.0	1	0
Total Count		13	12

Table 6. OCKMEx overlapping results

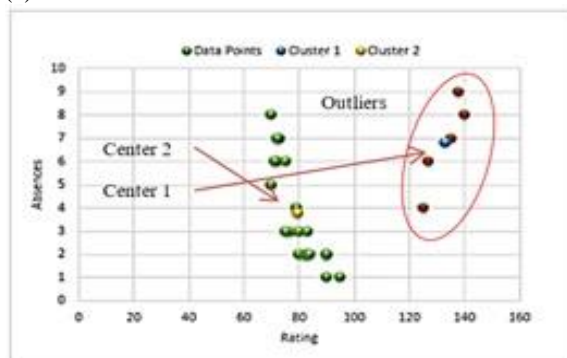
Vectors	Cluster Center	Cluster 1	Cluster 2
80, 2	90.0,2.0	0	0
90, 2	90.0,2.0	0	0
77, 3	73.0,6.0	0	0
70, 5	73.0,6.0	0	0
75, 3	73.0,6.0	0	1
72, 6	73.0,6.0	0	0
73, 7	73.0,6.0	0	0
80, 3	73.0,6.0	0	0
90, 2	90.0,2.0	0	0
79, 4	73.0,6.0	0	0
72, 7	73.0,6.0	0	0
71, 6	73.0,6.0	0	0
82, 2	90.0,2.0	0	0
83, 2	90.0,2.0	0	0
95, 1	90.0,2.0	0	0
90, 1	90.0,2.0	0	0
75, 6	73.0,6.0	0	1
70, 8	73.0,6.0	0	0
84, 2	73.0,6.0	0	0
83, 3	90.0,2.0	0	0
138, 9	90.0,2.0	0	0
135, 7	90.0,2.0	0	0
140, 8	90.0,2.0	0	0
125, 4	90.0,2.0	0	0
127, 8	90.0,2.0	0	0
Total Overlap Count		0	2



(a)



(b)



(c)

Figure 5. MCOKE simulation results with outliers for: (a) Iteration 1, (b) Iteration 2, and (c) Iteration 3

Tables 7 and 8 show the membership of each datapoint to its clusters. Based on the membership results, using the existing MCOKE, undiscovered data points overlap within clusters. Various studies stressed that the presence of outliers in the data samples would significantly affect the model in producing correct and accurate results [23]-[25]. With the influence of the outliers, MCOKE was unsuccessful in identifying data points with multi membership, making the algorithm ineffective in discovering valuable and vital data.

The summary results of the accuracy performance of the two experimentations performed using the synthetic data sample are shown in Table 9. Results indicate that the OCKMEx algorithm performed a higher accuracy rate of 100% in identifying data points that overlap even with outliers compared to the existing algorithm MCOKE. This proved that the OCKMEx algorithm is immune when outliers are mixed with actual values.

Table 7. MCOKE MT results with outliers

Vectors	Cluster Center	Cluster 1	Cluster 2
80, 2	79.55, 3.75	0	1
90, 2	79.55, 3.75	0	1
77, 3	79.55, 3.75	0	1
70, 5	79.55, 3.75	0	1
75, 3	79.55, 3.75	0	1
72, 6	79.55, 3.75	0	1
73, 7	79.55, 3.75	0	1
80, 3	79.55, 3.75	0	1
90, 2	79.55, 3.75	0	1
79, 4	79.55, 3.75	0	1
72, 7	79.55, 3.75	0	1
71, 6	79.55, 3.75	0	1
82, 2	79.55, 3.75	0	1
83, 2	79.55, 3.75	0	1
95, 1	79.55, 3.75	0	1
90, 1	79.55, 3.75	0	1
75, 6	79.55, 3.75	0	1
70, 8	79.55, 3.75	0	1
84, 2	79.55, 3.75	0	1
83, 3	79.55, 3.75	0	1
138, 9	133, 6.8	1	0
135, 7	133, 6.8	1	0
140, 8	133, 6.8	1	0
125, 4	133, 6.8	1	0
127, 8	133, 6.8	1	0
Total Count		5	20

Table 8. MCOKE overlapping results

Vectors	Cluster Center	Cluster 1	Cluster 2
80, 2	79.55, 3.75	0	1
90, 2	79.55, 3.75	0	1
77, 3	79.55, 3.75	0	1
70, 5	79.55, 3.75	0	1
75, 3	79.55, 3.75	0	1
72, 6	79.55, 3.75	0	1
73, 7	79.55, 3.75	0	1
80, 3	79.55, 3.75	0	1
90, 2	79.55, 3.75	0	1
79, 4	79.55, 3.75	0	1
72, 7	79.55, 3.75	0	1
71, 6	79.55, 3.75	0	1
82, 2	79.55, 3.75	0	1
83, 2	79.55, 3.75	0	1
95, 1	79.55, 3.75	0	1
90, 1	79.55, 3.75	0	1
75, 6	79.55, 3.75	0	1
70, 8	79.55, 3.75	0	1
84, 2	79.55, 3.75	0	1
83, 3	79.55, 3.75	0	1
138, 9	133, 6.8	1	0
135, 7	133, 6.8	1	0
140, 8	133, 6.8	1	0
125, 4	133, 6.8	1	0
127, 8	133, 6.8	1	0
Total Count		5	20

Table 9. Accuracy performance results

Test	Dataset	Algorithm	Cluster	Outlier	Overlap	Precision	Recall	F1-Measure
Experiment1	Synthetic	OCKMEx	2	0	2	1.0	1.0	1.0
		OCKMEx	2	5	2	1.0	1.0	1.0
Experiment2	Synthetic	MCOKE	2	5	0	0	0	0

4. CONCLUSION

In this study, we introduced a new overlapping algorithm called OCKMEx. This algorithm showed a better performance than the existing algorithms MCOKE for determining overlapping clusters and providing a more robust feature to outliers. Based on the results generated from experiments, OCKMEx provided a higher accuracy rate in identifying overlapping clusters even with outliers. An algorithm is a beneficial tool for clustering data objects and identifying overlapping clusters. Even with promising results, the researchers should do additional experiments and testing. In detail, a new calculation to isolate outliers is to be considered by the researchers as well. Additionally, the researchers suggested having another method considered a future study capable of excluding and separating the occurrence of outliers.




REFERENCES

- [1] H. Liu, X. Li, J. Li, and S. Zhang, "Efficient outlier detection for high-dimensional data," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2451–2461, 2018, doi: 10.1109/TSMC.2017.2718220.
- [2] N. Rehman, "Data mining techniques methods algorithms and tools," *International Journal of Computer Science and Mobile Computing*, vol. 6, no. 7, pp. 227-231, 2017.
- [3] T. F. G. Quilala, A. M. Sison, and R. P. Medina, "Securing electronic medical records using modified blowfish algorithm," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 6, no. 3, pp. 309-316, 2018, doi: 10.11591/ijeei.v6i3.493.
- [4] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: A review," *International Journal of Computer Applications*, vol. 120, no. 15, pp. 38-50, 2015, doi: 10.5120/21307-4126.




- [5] A. E. Danganan, A. M. Sison, and R. P. Medina, "OCA: Overlapping clustering application unsupervised approach for data analysis," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 14, no. 3, pp. 1471–1478, 2019, doi: 10.11591/ijeecs.v14.i3.pp1471-1478.
- [6] A. E. Danganan and E. De Los Reyes, "eHMCOKE: An enhanced overlapping clustering algorithm for data analysis," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 4, pp. 2212–2222, 2021, doi: 10.11591/eei.v10i4.2547.
- [7] R. Rooprekha and S. Perumal, "Data mining clustering techniques," *Engineering and Scientific International Journal*, vol. 3, no. 2, pp. 16–18, 2016.
- [8] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107964–108000, 2019, doi: 10.1109/ACCESS.2019.2932769.
- [9] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, Sept. 2014, doi: 10.1109/TKDE.2013.184.
- [10] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagão, "Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 954–960, doi: 10.1109/ICMLA.2016.0172.
- [11] U. Porwal and S. Mukund, "Credit card fraud detection in e-commerce: An outlier detection approach," *arXiv*, 2018, doi: 10.48550/arXiv.1811.02196.
- [12] K. Alrawashdeh and C. Purdy, "Toward an online anomaly intrusion detection system based on deep learning," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 195–200, doi: 10.1109/ICMLA.2016.0040.
- [13] G. B. Gebremeskel, C. Yi, Z. He, and D. Haile, "Combined data mining techniques based patient data outlier detection for healthcare safety," *International Journal of Intelligent Computing and Cybernetics*, vol. 9, no. 1, pp. 42–68, 2016, doi: 10.1108/IJICC-07-2015-0024.
- [14] K. Singh and S. Upadhyaya, "Outlier detection: Applications and techniques.," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 307–323, 2012.
- [15] D. Feldman and L. J. Schulman, "Data reduction for weighted and outlier-resistant clustering," *Proceedings of the 2012 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2012, pp. 1343–1354, doi: 10.1137/1.9781611973099.106.
- [16] C. Whelan, G. Harrell, and J. Wang, "Understanding the k-medians problem," *Proceedings of the International Conference on Scientific Computing (CSC)*, 2015, pp. 219–222.
- [17] B. Shathya, "Predicting students' performance using k-median clustering," *International Journal of Data Mining Techniques and Applications (IJDMTA)*, vol. 4, no. 2, pp. 67–69, 2015, doi: 10.20894/ijdmata.102.004.002.004.
- [18] S. Baadel, F. Thabtah, and J. Lu, "Overlapping clustering: A review," *2016 SAI Computing Conference (SAI)*, 2016, pp. 233–237, doi: 10.1109/SAI.2016.7555988.
- [19] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. J. Mooney, "Model-based overlapping clustering," *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 532–537, doi: 10.1145/1081870.1081932.
- [20] C. Fayet, A. Delhay, D. Lolive, and P.-F. Marteau, "Unsupervised classification of speaker profiles as a point anomaly detection task," *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 2017, pp. 152–163.
- [21] K. Limthong, "Real-time computer network anomaly detection using machine learning techniques," *Journal of Advances in Computer Networks*, vol. 1, no. 1, pp. 1–5, 2013, doi: 10.7763/jacn.2013.v1.1.
- [22] S. Baadel, F. Thabtah, and J. Lu, "MCOKE: Multi-cluster overlapping k-means extension algorithm," *International Journal of Computer, Control, Quantum and Information Engineering*, vol. 9, no. 2, pp. 427–430, 2015.
- [23] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "A multi-step outlier-based anomaly detection approach to network-wide traffic," *Information Sciences*, vol. 348, pp. 243–271, 2016, doi: 10.1016/j.ins.2016.02.023.
- [24] A. Barai (Deb) and L. Dey, "Outlier detection and removal algorithm in k-means and hierarchical clustering," *World Journal of Computer Application and Technology*, vol. 5, no. 2, pp. 24–29, 2017, doi: 10.13189/wjcat.2017.050202.
- [25] D. Pahuja and R. Yadav, "Outlier detection for different applications: review," *International Journal of Engineering Research and Technology*, vol. 2, no. 3, pp. 1–8, 2013.

BIOGRAPHIES OF AUTHORS



Dr. Alvincent E. Danganan    has been a faculty member of Tarlac State University College of Computer Studies, Philippines since 2003. He served as the Chairperson of the Department of Information Systems from 2013 to 2016. He was also designated as the College Extension Service Chairperson for the same period and has conducted projects and presentations awarded at the institutional level. He also served as the Chairperson of the Computer Science Department from 2019–2020. Currently he is the Dean of Tarlac State University College of Computer Studies. He is also one of the area coordinators of the Philippines Society of Information Technology Educators Central Luzon, Philippines chapter. His research interest includes data mining, machine learning and data analytics. He has authored publications on data mining in Scopus-indexed journals. The author may be reached at avdanganan@tsu.edu.ph.



Ms. Regina P. Arceo    is currently an IT Instructor at the College of Computer Studies, Tarlac State University. Presently, she is enrolled at the graduate program of the university (Master's in information technology) and has already earned the required academic units leading to a master's degree in same specialization. Among her other accomplishments, Ms. Arceo has been also engaged on IT-related webinars, trainings, and conferences both national and international that have helped her grow professionally. One of the International conferences she attended was the "Global Young Scientist Summit Singapore 2021" which is an annual international summit held every January and which was endorsed by DOST where 20 nominees around Philippines were selected to participate. She can be contacted at email: rparceo@tsu.edu.ph.