Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 4, August 2021, pp. 2212~2222
ISSN: 2302-9285, DOI: 10.11591/eei.v10i4.2547

eHMCOKE: an enhanced overlapping clustering algorithm for data analysis

Alvincent E. Danganan, Edjie De Los Reyes
Tarlac State University, Romulo Blvd. San Vicente, Tarlac City, Philippines

Article Info

Article history:

Received Apr 17, 2020
Revised May 20, 2021
Accepted Jun 15, 2021

Keywords:

K-means
MAD
Maxdist
Outliers
Overlap

ABSTRACT

Improved multi-cluster overlapping k-means extension (IMCOKE) uses median absolute deviation (MAD) in detecting outliers in datasets makes the algorithm more effective with regards to overlapping clustering. Nevertheless, analysis of the applied MAD positioning was not considered. In this paper, the incorporation of MAD used to detect outliers in the datasets was analyzed to determine the appropriate position in identifying the outlier before applying it in the clustering application. And the assumption of the study was the size of the cluster and cluster that are close to each other can led to a higher runtime performance in terms of overlapping clusters. Therefore, additional parameters such as radius of clusters and distance between clusters are added measurements in the algorithm procedures. Evaluation was done through experimentations using synthetic and real datasets. The performance of the eHMCOKE was evaluated via F1-measure criterion, speed and percentage of improvement. Evaluation results revealed that the eHMCOKE takes less time to discover overlap clusters with an improvement rate of 22% and achieved the best performance of 91.5% accuracy rate via F1-measure in identifying overlapping clusters over the IMCOKE algorithm. These results proved that the eHMCOKE significantly outruns the IMCOKE algorithm on mosts of the test conducted.

Corresponding Author:

Alvincent E. Danganan
College of Computer Studies
Tarlac State University
Romulo Blvd. San Vicente, Tarlac City, Philippines
Email: avdanganan@tsu.edu.ph, avdanganan836@gmail.com

1. INTRODUCTION

Extraction of patterns from data is a method called data mining [1]. In knowledge discovery in database (KDD) process, data mining is an important part that is used to find significant information and discover hidden patterns from the huge collection of data [2]. Mining is used to dig through data and discover new knwoledge from a various information which is then used in many applications which is sometimes referred to us data science [3]. Preventive medicine is one of the fields which uses knowledge discovery in data to analyze partient information for diagnosis of the diseases. There are two categories of functions involved in data mining: supervised and unsupervised learning [4]. In supervised learning, the model is trained on a labeled data sets while unsupervised learning the model is used to identify patterns in unlabeled data sets [5].

Clustering can be considered as an unsupervised learning technique. It is one of the most significant and challenging data mining techniques in the knowledge discovery process. The goal of clustering is to discover groups of objects from unlabeled data such that all similar data object is within the same clusters while dissimilar data object from different clusters [6].

However, most of the real world data sets have overlapping information [7] where data objects or patterns can belong to one or more clusters. Numerous research works have focused on this problem known as overlapping clustering technique. for example, in a social network, a person may belong to two or more communities [8]. In music, emotion data set can be categorized as relaxing and happy at the same time [9]. A method called scalable spectral clustering was used to detect underlying communities in a larger networks [10]. The multi-cluster overlapping k-means or MCOKE extension was newly introduced as another method in segmenting a data into clusters as well as finding overlapped data [11]. Despite providing better results in detecting data that overlapped, MCOKE is sensitive to outlier which can have negative effects on the accuracy in identifying overlapping objects within clusters. Therefore, improvement of MCOKE algorithm have been introduced for better performance in identifying overlap clusters. According to recent studies, [12]-[14] observation of unusual value has a significant role in the field of data mining. The IMCOKE [15] algorithm was presented that focuses on the incorporation of median absolute deviation (MAD) as the outlier detection method used to detect outliers. However, the study did not consider the positioning of MAD procedure applied in the algorithm. Furthermore, the concentration of IMCOKE algorithm is only on the measurement between the distance of the data to the centroid in finding the data that overlap within clusters and disregard other vital parameters.

In this paper, the study is to enhance the algorithm by determining the best position of MAD in identifying the outlier before applying it in the algorithm procedures. The study will examine if the outlier detection positioning affects its capability in detecting outliers in the datasets. In addtion, measurement of parameters such as distance between clusters and radius of the clusters are also considered in the study to achieve faster and more accurate identification of overlapping clusters.

2. RESEARCH METHOD

2.1. Outlier detection

In the process of detecting the outliers, each data objects were collated and classified inascending order. To detect the anomaly in the data, first compute the median value (M_i), where Mi is the median of the sequence of distances of data objects. Then, compute the MAD values by deducting the median from each distance of a data object. Next, the calculated MAD values were classify in ascending order, and the median of absolute deviation values were determined. After this, the median was multiplied by b, the contrast b equal to 1.4826 which is constant linked to the assumption of normality of the data [16]. The (1) shows the MAD formula.

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

Once MAD is calculated, a threshold value was determined which serve as a basis to guide the outlier detection. A study [17] suggest that the values of 3, 2.5, and 2 as the threshold value of an outlier. A decision value was computed using (2). Values greater than or smaller than the decision value are considered outliers which are removed from the clusters. In this study, a threshold value of 2.5 was adopted since it provides a reasonable choice for outlier detection [18].

$$\text{Decision Value}=\text{M} \pm \text{threshold value} * \text{MAD} \quad (2)$$

This method will be terminated once all outliers have been isolated from the data sets.

2.2. Radius of cluster and distance between cluster calculation

Radius of a cluster and the distance between clusters [19] are two measurements that were considered to improve the algorithm in terms of time spent in identifying overlap clusters. To get the radius, R, of the cluster, the mean distance of the data in the cluster is multiplied with the number of clusters as defined in (3). Illustration for the calculation is depicted in Figure 1.

$$\mathbf{R} = \frac{\text{add all distance of pints from the centroid}}{\text{number of points}} * \text{number of clusters} \quad (3)$$

To obtain the distance between clusters, D, (4) is used. A sample calculation is shown in Figure 2.

$$\boldsymbol{D} = \frac{\text{add all distance of points from the centroid}}{\text{number of points}} * \text{no\# of clusters} \quad (4)$$

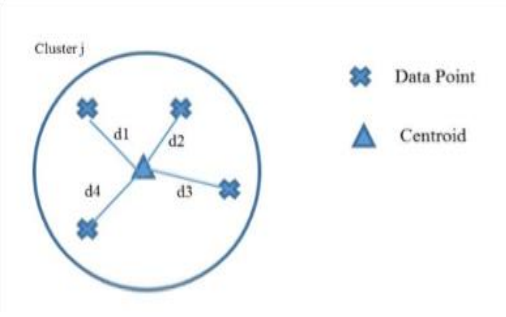

Figure 1. Sample calculation of a cluster radius

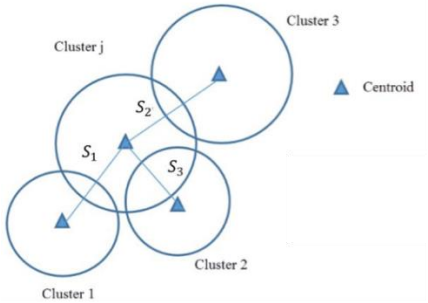

Figure 2. Sample calculation for distance b/w clusters

2.3. Enhanced MCOKE algorithm

Two strategies were employed to the algorithm. The first strategy was to remove outliers. The second strategy involved the incorporation of added parameters. The formation of the new and Enhanced MCOKE algorithm is shown in Figure 3.

The enhanced MCOKE (eHMCOKE) algorithm consists of three phases. Phase 1 is the used of MAD to discover the anomalous value (outlier) in the datasets and this value is isolated before the clustering of data. Phase 2 is to group the data into cluster using K-means algorithm. Then finally in the last Phase, overlapped clusters were identified with the used of maxdist and the added parameters modifying the previous procedure that identify clusters that overlapped.

Input: Number of k Clusters Centroid and Data Points

Output: Membership Table (MT)

Phase 1

- Get calculated distance of object on each cluster
- Rank distance in ascending order
- Get median value Mi
- Calculate the Median absolute deviation
- Multiply MAD by b, where b=1.4826
- Calculate the decision value
- Compare the calculated distance in a cluster with the decision value
- If distance is greater or less than decision value outlier is detected
- Remove outlier

Phase 2

- Randomly select initial cluster centroid
- Calculate the distance between each object and cluster centroid using Euclidian distance.
- Assign objects to its nearest cluster centroid
- Re-calculate cluster centroid
- If not converged, repeat from step 2.
- Save maximum distance of object allowed in a cluster.

Phase 3

- Calculate radius and distance of cluster
- Compare distance of object of cluster with high probability to overlap with another final centroid.
- Assign objects to another cluster if distance is less than maxdist.

Figure 3. eHMOCKE algorithm

2.4. Evaluation

The performance of the eHMCOKE was evaluated based on its accuracy and speed. This process allows for the comparison between the IMCOKE and the eHMCOKE algorithm and determines whether one algorithm outperform or superior to another one.

a. Speed or execution

The speed was measured by subtracting the elapsed time from the start time.

b. Percentage of improvement

The percentage improvement was computed to compare the performance of the eHMCOKE and the IMCOKE algorithms (5).

$$\text{PI} = \frac{\text{I}-\text{O}}{\text{O}} * 100 \quad (5)$$

2.5. Accuracy

Recall, precision and F-measure were calculated over pairs of points used in the evaluation of the accuracy of overlapping clustering results. Precision is calculated based on the correct identification of pairs in the same cluster and recall is the actual pairs that were identified. The formula for precision is shown in (6) while that for recall is shown in (7) [20].

$$\text{Precision} = \frac{\text{Number of Correctly Identified LinkeD Paird}}{\text{Number of Identified Linked Pairs}} \quad (6)$$

$$\text{Recall} = \frac{\text{Number of Correclty Identified Linked Pairs}}{\text{Number of True Linked Pairs}} \quad (7)$$

The actual calculation for precision and recall were made by using true outliers as few false positives. A large number of false positives indicates a low precision. A recall is to measure the performance of the outlier detection in capturing the most or all outliers as few false negatives as possible. A low recall indicates a large number of false negatives. The (8) shows the formula for precision while the (9) shows the formula for recall [21].

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP +FN}} \quad (9)$$

Where true positives (TP) is the accurately predicted true outliers, false positives (FP) is the predicted true outlier, but is not, and false negative (FN) is the predicted not an outlier, but it is a true outlier.

To model the desired precision and recall, the F-measure, also referred to as the F1 score combined with precision and recall was used. F1 score computes the weighted harmonic mean of recall and precision [22]. Having higher F1 score result constitute to an excellent detection accuracy, where 0 mean the worst and 1 mean the perfect detection [23]. The (10) shows the calculation of F-measure.

$$\text{F}_1 \text{ Score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (10)$$

3. RESULTS AND DISCUSSION

In this section, three experimentations were conducted to test the eHMCOKE algorithm. Synthetic and Real datasets were used.

3.1. Experiment 1

The objective of the first experiment was to contrast the results between the two strategies for outlier detection used in the clustering analysis. The study intended to compare the accuracy rate of the outlier detection procedure MAD before clustering of data and after the clustering of data. Experiments were made on synthetic and real data sets.

Two attributes (Rating, Absences) with 50 instances are form in the synthetic data set. Five outliers were intentionaly incorporated in the data set; therefore, 45 instances are normal, and five instances are unusual data or also known as outliers (Student 46 to Student 50). Table 1 shows the synthetic dataset.

Table 1. Synthetic experimental datasets

STUDENT	Rating	Absences	STUDENT	Rating	Absences
Student 1	80	3	Student 26	75	4
Student 2	90	2	Student 27	72	6
Student 3	77	3	Student 28	84	2
Student 4	70	5	Student 29	83	3
Student 5	78	3	Student 30	82	2
Student 6	72	6	Student 31	90	4
Student 7	73	7	Student 32	98	3
Student 8	80	3	Student 33	99	2
Student 9	90	2	Student 34	90	4
Student 10	79	4	Student 35	72	5
Student 11	72	7	Student 36	77	3
Student 12	71	6	Student 37	76	4
Student 13	82	2	Student 38	74	6
Student 14	83	2	Student 39	72	3
Student 15	95	1	Student 40	68	8
Student 16	90	1	Student 41	65	9
Student 17	74	6	Student 42	64	9
Student 18	70	8	Student 43	63	9
Student 19	80	6	Student 44	62	5
Student 20	78	7	Student 45	71	4
Student 21	78	3	Student 46	150	9
Student 22	70	5	Student 47	155	7
Student 23	88	2	Student 48	140	8
Student 24	90	2	Student 49	120	4
Student 25	100	1	Student 50	135	5

In this work, synthetic dataset was used for the first experimental run, data were plotted through 2-dimensional spaces as shown in Figure 5. Then, the outlier detection MAD procedure was tested to find outliers before the clustering of the data. Figure 6 shows the visualization results, red dots are the outliers found in the dataset recognized by MAD before performing the clustering method. Found outliers were removed from the datasets.

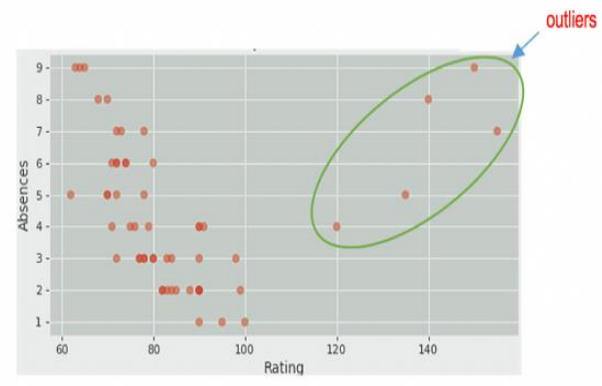

Figure 4. Scatter plot of synthetic data set

Figure 5. Simulation result of detected outliers before the clustering of data

The same synthetic dataset was processed for the identification of outlier. This time, MAD was tested after the clustering of data. First, data objects were segregated into various of clusters with the used of K-means algorithm. K was initiated randomly, then cluster centroids were formed based on the initial number of K where data objects are being assigned. For this experiment, the user selects three (3) where K=3 clusters centroid and based on its Euclidian distance measurement each data was assigned to its nearest cluster. The test data was run five (5) to 20 times with a dissimilar k number of clusters, and the best result was used in the experiment. As shown in Figure 7, the output of 50 data objects with 2 clusters.

The second experiment was conducted to test the outlier detection MAD on real datasets obtained from UCI machine repository. In this experiment, Iris plant dataset was considered. The Iris plant dataset

contains 155 instances with two attributes, and five are considered outliers. Results are shown in Figure 8 and Figure 9. The results of the tests conducted are summarized in Table 2.

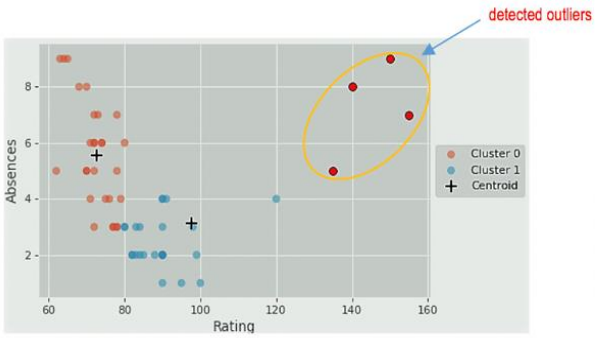

Figure 6. Simulation result of detected outliers after the clustering of data

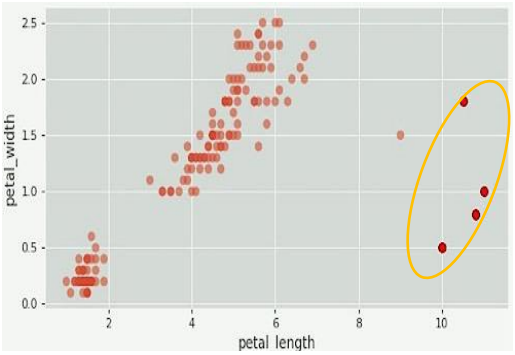

Figure 7. Simulation result of detected outliers before the clustering of data

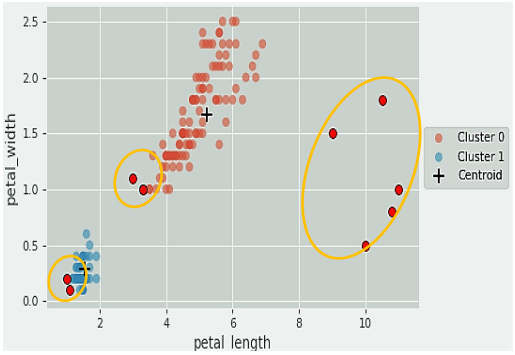

Figure 8. Simulation result of detected outliers after the clustering of data

Table 2. Accuracy results

Datasets	Outlier detection applied	Precision	Recall	F-measure
Synthetic	before clustering	1.0	1.0	**1.0**
	after Clustering	1.0	0.80	0.89
Iris Plants	before clustering	1.0	0.80	**0.89**
	after Clustering	0.56	1.0	0.71

Based on the results, MAD achieved a higher accuracy rate of 100% before the clustering of data under synthetic dataset. For the iris plants dataset, MAD obtained the best performance of 89% accuracy rate before the clustering of data.

As seen in Table 2, the implementation of MAD before clustering of data achieved higher performance accuracy rate in terms of finding outliers in the datasets. The outcomes of this series of experiments gave a piece of substantial evidence that the detection of outlier before performing clustering analysis works well with different types of datasets.

3.2. Experiment 2

The aim of the second experiment is to test whether the additional parameters added in the algorithm significantly affects the time to detect objects that overlaps.

To test the runtime execution of each algorithm, synthetic datasets were used considering two Gaussian clusters datasets (G2-2-30, G2-2-50), one high dimensional dataset and one compound dataset [24]. To obtain a clear insight of the clustering capability of different clustering methods, a simulation of the clustering results on each dataset was done. The simulation results for different scenarios using the IMCOKE and eHMCOKE are shown in Figure 10. Summary of the experimental results for runtime execution of the two algorithms is shown in Table 3.

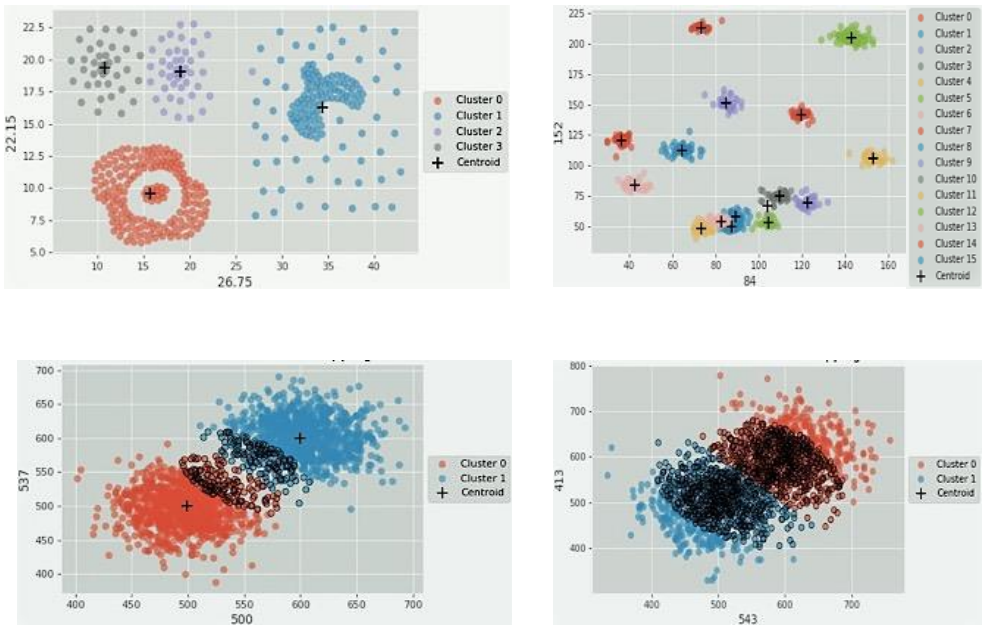

Figure 10. Simulation results

Table 3. Runtime performance analysis

Datasets	No. of objects	No. of clusters	Execution Time (in Seconds)	
			IMCOKE	eHMCOKE
Compound	398	4	0.05	**0.02**
High dimensional	1024	16	0.62	**0.08**
G2-2-30	2048	2	0.10	**0.05**
G2-2-50	2048	2	0.17	**0.05**

Results indicate that the measurements for cluster size and distances between clusters affect the execution time in identifying objects that overlap between clusters. The IMCOKE ignores this size of the clusters and distance between clusters which makes the identification of overlap clusters quite time-consuming especially on a more significant number of clusters. This makes the eHMCOKE perform better in terms of runtime execution even with a profoundly more substantial amount of data objects.

3.3. Experiment 3

The third experiment is to test the accuracy performance of the two algorithms in terms of identifying overlap clusters, the synthetic data set (Synthetic 1) used is composed of 37 observations with two attributes, three considered as linked pairs that will overlap, and two treated as outliers. Figure 11 illustrate the simulation result of the actual data.

The study performed three tests with two approaches, one with the used of the IMCOKE algorithm and another with eHMCOKE algorithm for comparison. In the IMCOKE algorithm, segmentation of objects into clusters was established first before the detection of outliers or before the incorporation of MAD. In this experiment, the user inputted two K clusters centroid, and clusters are formed once each object is assigned to its nearest cluster center.

Based on the simulation result, IMCOKE consider the outliers as members of one cluster therefore outliers were not been identified because clusters are formed prior to the identification of outliers. Then maxdist was used to identify the belonging of objects to multiple clusters. As shown in Figure 12, using the IMCOKE algorithm, there are no identified overlaps.

In the eHMCOKE algorithm experiment, the study considered the incorporation of MAD before the clustering of dataset since it results a higher accuracy rate in detecting outliers based on the first experimentation that was conducted. The same synthetic data set (Syntheic 1) was used. Before segmenting the data to its assign cluster, the objective of eHMCOKE is to isolate the outliers in the datasets with the used of MAD. With the integration of MAD as shown in Figure 13 evidently display that outlier were accurately discovered.

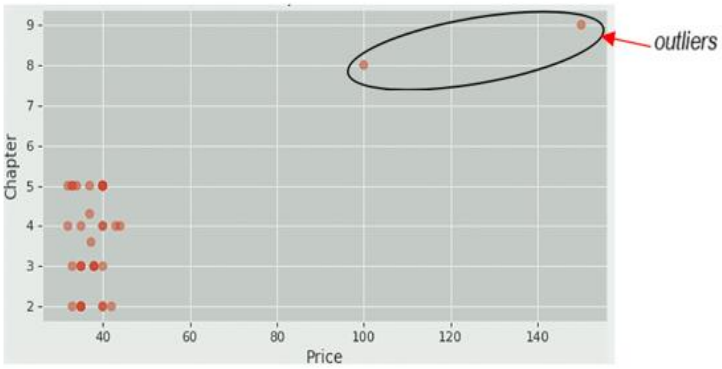

Figure 11. Scatter plot of synthetic data set

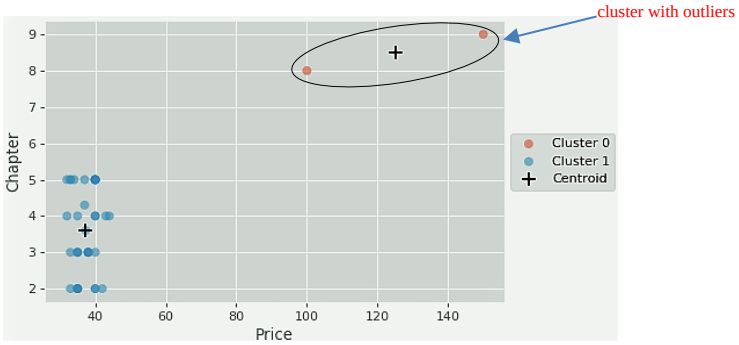

Figure 12. Non-identification of overlapping clusters

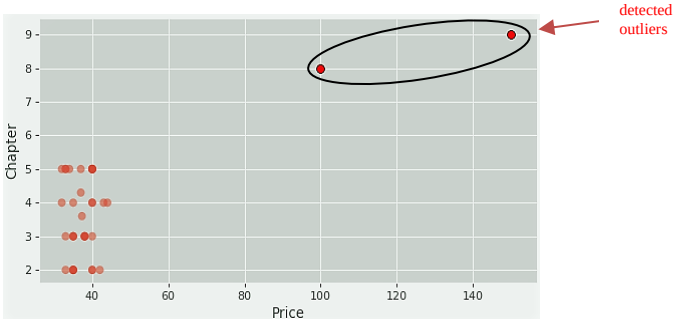

Figure 13. Outliers in the data set

Researchers emphasized that isolating unusual data in the dataset produce a more correct and precise outcome in the field of data mining thus isolating of this data from the dataset is significant [25], [26]. These found outliers are separated from the normal dataset and were no longer considered part of the procedure in detecting clusters that overlap. Figure 14 shows the visualization result of a cleaned dataset.

The same dataset was processed, the algorithm takes an input of two clusters centroid to form a cluster. Followed by the identification of overlap clusters. In this stage, additional parameters such as the measurement of radius and distance between clusters were added into the algorithm procedure. The study assumed that these parameters could also assist in the overlapping clustering processes. Calculating these parameters followed by the used of maxdist will have a high probability in finding patterns that overlap with other clusters. As shown in Figure 16, the simulation results of the eHMCOKE proved that the enhance algorithm was able to accurately detect the three considered linked pairs that overlap in the dataset.

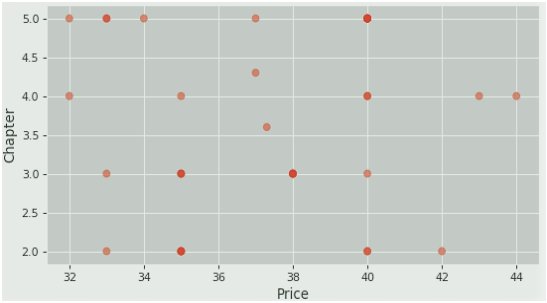

Figure 14. Patterns without outliers

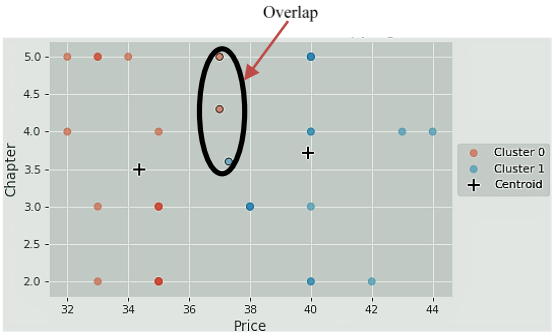

Figure 15. Overlapping results of the enhanced MCOKE

A further test was conducted between the eHMCOKE algorithm and the IMCOKE algorithm under a larger scale of data. The study used Gaussian synthetic dataset (Synthetic 2); it was composed of 2048 observations with 332 overlap data. Figure 16 shows the simulation result of the actual data. The test data contains two clusters, and the results are shown in Figure 17 and Figure 18.

The summary of the experimental results for all the cluster combinations performed using the two synthetic datasets are shown in Table 4. Based on the results, the eHMCOKE achieved the best performance of 100% under Synthetic 1 dataset, which means that the eHMCOKE algorithm outruns the IMCOKE algorithm. For the Synthetic 2 dataset, the eHMCOKE algorithm obtained a higher accuracy rate of 83% which outperformed the IMCOKE algorithm. Table 4 shows that the eHMCOKE achieved higher performance accuracy rate in terms of finding overlap data.

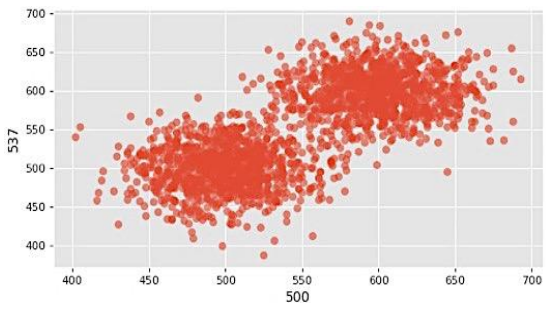

Figure 16. Scatter plot under gaussian data set

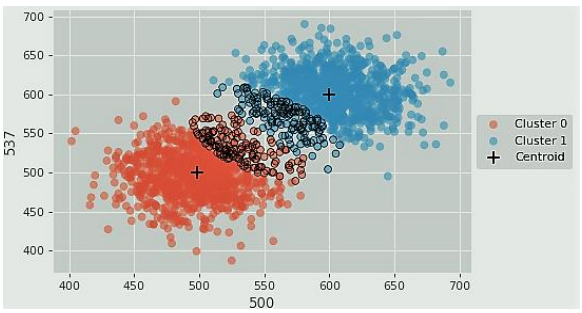

Figure 17. Overlap result of the original MCOKE

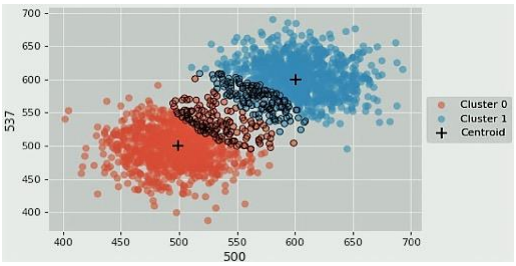

Figure 18. Overlap result of the enhanced MCOKE

Table 4. Overlapping clustering result

Dataset	Algorithm	No. of Clusters	Overlap	Precision	Recall	F1-Measure
Synthetic 1	IMCOKE	2	0	0.0	0.0	0.0
	eHMCOKE		3	1.0	1.0	**1.0**
Synthetic 2	IMCOKE	2	321	1.0	0.69	0.82
	eHMCOKE		324	1.0	0.70	**0.83**

4. CONCLUSIONS AND RECOMMENDATIONS

Based on the findings of this research, MAD procedure was applied before clustering of the data in eHMCOKE since it results in consistently higher accuracy rate compared to the application of MAD after clustering of the datasets. The eHMCOKE algorithm performed faster over the IMCOKE algorithm with an improvement rate of 22% in identifying overlapping clusters. The eHMCOKE algorithm achieved an improvement rate of 99% over the IMCOKE algorithm based on its F1-score. The conclusions stated above shows that the incorporation of outlier detection prior to clustering improves the performance of the eHCOKE to detect outliers. This has led to better identification of overlap clusters. The used of the additional parameters also contributed to the enhancement of the algorithm in terms of runtime execution. Thus, the study has successfully achieved its objective of producing an eHMCOKE algorithm with better performance compared to the existing IMCOKE algorithm.

Furthermore, it is recommended that other test measures such as FBCubed and Pair-based evaluation may be considered to evaluate the performance of the Enhanced algorithm. Since the eHMCOKE still uses the traditional k-means algorithm, it is still sensitive to the random initialization of the cluster's centroid. An alternative approach to the random initialization is recommended. eHMCOKE can only be used with numeric data input, improvement of the algorithm may be done for it to accept textual inputs. New applications of the enhanced algorithm may be exhausted.

REFERENCES

[1] S. Vijayarani and S. Nithya, "An efficient clustering algorithm for outlier detection," *Int. J. Comput. Appl.*, vol. 32, no. 7, pp. 22-27, 2011.

[2] P. Kaur and K. Kaur, "A Review on Outlier Detection for Data," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 4, no. 7, pp. 14373-14376, 2016, doi: 10.15680/IJIRCCE.2016.0407176.

[3] C. Virmani, A. Pillai, and D. Juneja, "Clustering in aggregated user profiles across multiple social networks," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 6, pp. 3692-3699, 2017, doi: 10.11591/ijece.v7i6.pp3692-3699.

[4] 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.

[5] V. M. A. Souza, R. G. Rossi, G. E. A. P. A. Batista, and S. O. Rezende, "Unsupervised active learning techniques for labeling training sets: An experimental evaluation on sequential data," *Intell. Data Anal.*, vol. 21, no. 5, pp. 1061-1095, 2017, doi: 10.3233/IDA-163075.

[6] 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, doi: 10.5220/0004830002790286.

[7] 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, doi: 10.1016/j.eswa.2016.09.025.

[8] S. Jiang, X. Lu, C. Xie, and S. Cai, "Adaptive finite-time control for overlapping cluster synchronization in coupled complex networks," *Neurocomputing*, vol. 266, pp. 188-195, 2017, doi: 10.1016/j.neucom.2017.05.031.

[9] C. Ben n'cir, E. Nadia, G. Cleuziou *Overview of Overlapping Partitional Clustering Algorithms*, no. January. Cham: Springer International Publishing, 2015.

[10] H. Van Lierde, G. S. Member, and T. W. S. Chow, "Scalable Spectral Clustering for Overlapping Community Detection in Large-Scale Networks," *IEEE Trans. Knowl. Data Eng.*, vol. PP, no. c, p. 1, 2019, doi: 10.1109/TKDE.2019.2892096.

[11] S. Baadel, F. Thabtah, and J. Lu, "MCOKE : Multi-Cluster Overlapping K-Means Extension Algorithm," vol. 9, no. 2, pp. 427-430, 2015.
[12] K. Yan, X. You, X. Ji, G. Yin, and F. Yang, "A Hybrid Outlier Detection Method for Health Care Big Data," *2016 IEEE Int. Conf. Big Data Cloud Comput. (BDCloud), Soc. Comput. Netw. (SocialCom), Sustain. Comput. Commun.*, pp. 157-162, 2016, doi: 10.1109/BDCloud-SocialCom-SustainCom.2016.34.
[13] K. Singh and S. Upadhyaya, "Outlier Detection: Applications And Techniques.," *Int. J. Comput. …*, vol. 9, no. 1, pp. 307-323, 2012.
[14] H. Liu, X. Li, J. Li, and S. Zhang, "Efficient Outlier Detection for High-Dimensional Data," *IEEE Trans. Syst. Man, Cybern. Syst.*, pp. 1-11, 2017, doi: 10.1109/TSMC.2017.2718220.
[15] 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, doi: 10.11591/ijeei.v6i4.499.
[16] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *J. Am. Stat. Assoc.*, vol. 88, no. 424, pp. 1273-1283, 1993, doi: 10.1080/01621459.1993.10476408.
[17] J. Miller, "Short Report: Reaction Time Analysis with Outlier Exclusion: Bias Varies with Sample Size," *Q. J. Exp. Psychol. Sect. A*, vol. 43, no. 4, pp. 907-912, 1991, doi: 10.1080/14640749108400962.
[18] 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, doi: 10.1016/j.jesp.2013.03.013.
[19] T. Limungkura and P. Vateekul, "Partition-based Overlapping Clustering Using Cluster ' s Parameters and Relations," *IEEE*, pp. 144-149, 2017.
[20] A. Banerjee, C. Krumpelman, and R. J. Mooney, "Model-based overlapping clustering Model-based Overlapping Clustering," in *SIGKD ACM International Conference on Knowledge Discovery*, 2014, no. January 2005, doi: 10.1145/1081870.1081932.
[21] H. Jawed, Z. Ziad, M. M. Khan, and M. Asrar, "Anomaly detection through keystroke and tap dynamics implemented via machine learning algorithms," *Turk J Elec Eng Comp Sci*, pp. 1698-1709, 2018, doi: 10.3906/elk-1711-410.
[22] C. Fayet and D. Lolive, "Unsupervised Classification of Speaker Profiles as a Point Anomaly Detection Task," in *Proceeding of Machine Learning Research*, 2017, pp. 152-163.
[23] K. Limthong, "Real-Time Computer Network Anomaly Detection Using Machine Learning Techniques," *J. Adv. Comput. Networks*, vol. 1, no. 1, pp. 1-5, 2013, doi: 10.7763/JACN.2013.V1.1.
[24] P. Fr and S. Sieranoja, "K-means properties on six clustering benchmark datasets Pasi Fr anti," *Appl. Intell.*, pp. 55-57, 2018.
[25] A. Barai (Deb) and L. Dey, "Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering," *World J. Comput. Appl. Technol.*, vol. 5, no. 2, pp. 24-29, 2017, doi: 10.13189/wjcat.2017.050202.
[26] P. K. Sharma, H. Haleem, and T. Ahmad, "Improving Classification by Outlier Detection and Removal," in *49th Annual Convention of the Computer Society of India CSI*, 2015, no. December 2014, doi: 10.1007/978-3-319-13731-5.

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 which have been 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 Technolody Educators Central Luzon, Philippines chapter. His research interest includes data mining, machine learning and Artificial Intelligence. He has authored publications on data mining in Scopus-indexed journals. The author may be reached at avdangana@tsu.edu.ph.

Edjie M. De Los Reyes is an Assistant Professor in Tarlac State University with 19 years of academic experience. At present, he is designated as the Research Director of the university. Likewise, he has served as an Associate Dean from 2014 to 2016. He has also earned numerous skills certifications such as Cisco Certified Network Associate, Cisco Certified Academic Instructor, Microsoft Office Specialist, and Electronic Data Processing Specialist - Programmer. Moreover, he is an active member of different academic and research organizations like Philippine Society of Information Technology Educators (PSITE), Philippine Schools Universities and Colleges Computer Education and Systems Society (PSUCCESS), International Association of Multidisciplinary Research to name a few. He has also several published research papers focused on data security in Scopus-indexed journals. The author may be reached through emdelosreyes@tsu.edu.ph