International Multilingual Journal of Science and Technology (IMJST)
ISSN: 2528-9810
Vol. 9 Issue 3, March - 2024

Determination Of Optimal Number Of Clusters Using Gap Statistics And Elbow Methods

Afolayan Jimoh Jacob [1]
Department of Electrical/Electronics Engineering,
University of Uyo, Akwa Ibom State, Nigeria
afolayan.jimoh@yahoo.com

Ikpe, Joseph Daniel[2]
Department of Electrical/ Electronic Engineering
Akwa Ibom State Polytechnic Ikot Osurua , Ikot Ekepene

Chikezie Samuel Aneke[3]
Department Of Computer Engineering
University Of Uyo, Akwa Ibom State

***Abstract*— In this paper, determination of optimal number of Clusters using Gap Statistics and Elbow methods is presented. The algorithms to determine optimal number of cluster using Gap statistics and the Elbow methods are presented along with some of the key mathematical models associated with each of the two methods. A program developed using Python 3 in Pycharm development environment was used to simulate the Gap Statistics and Elbow algorithms. The case study considered has an area, A with dimensions of 800m × 800m which amounts to a 640 km^2 area. The results obtained for the Elbow method is presented in Figure 2. The results show that the Elbow method gave optimal number of clusters as four, hence the 5000 sensors nodes are grouped into four clusters given as; $Cluster\ 0$, $Cluster\ 1$, $Cluster\ 2$ and $Cluster\ 3$. On the other hand, the results show that the Gap Statistics method gave optimal number of clusters as five, hence the 5000 sensors nodes are grouped into five clusters given as; $Cluster\ 0$, $Cluster\ 1$, $Cluster\ 2$, $Cluster\ 3$ and $Cluster\ 4$. In addition, the results show that it took 5.8 seconds for the Elbow method to determine the optimal number of clusters whereas it took the Gap statistics method 5.3 seconds. There is therefore, about 0.5 seconds gained by using the Gap statistics method which is about 8.62069 % improvement in the implementation time over the Elbow method. Therefore, the Gap Statistics method is recommended given its better implementation time.**

***Keywords*— Optimal Number of Clusters, Gap Statistics, Sensor Node, Elbow Method, Clustering Algorithms**

1. INTRODUCTION

Over the years, wireless sensors have been deployed as part of diverse systems [1,2,3,4]. The advancement in the sensor technologies, computational technologies and communication technologies have given rise to more applications of these sensors in more systems like smart cities, smart health, smart transport and other areas that most often require deployment of large numbers of sensor nodes [5,6,7,8]. Generally, wireless sensors are known to be resource constrained. Many of the sensors are battery powered with limited lifespan. As such, deployment of such sensors requires careful planning to maximize the battery lifespan [9,10,11,12,13]. Clustering has been used as one of the means of achieving such energy management in wireless sensor network [14,15,16].

In clustering, the sensors in the network are grouped into clusters. However, the number of cluster suitable for a given network must be determined [17,18,19]. This is achieved using computational techniques. In this work, gap statistics and elbow methods are considered [20,21,22]. The two methods are considered because of their simplicity and accuracy in computing the optimal number of clusters. Moreover, the two methods approaches the problem using different concepts which do affect their solutions and computation time. As such, in this work, the algorithm for the two methods are presented and their performances are evaluated through simulations based on certain number of sensor nodes and network coverage area. In all, the study seek to identify the more accurate and more efficient method to be recommended for cluttering applications in wireless sensor network design.

2. METHODOLOGY

2.1 DETERMINATION OF OPTIMAL NUMBER OF CLUSTERS USING GAP STAT

In the Gap Statistics (GS) approach, first a null reference data distribution is given and it is used as the reference value to compare the cluster congestion of compactness. The optimum cluster number is achieved at the point at which the congestion value is the highest with respect to the reference curve. The point at which the congestion is highest, from the reference curve is considered as the optimal number of clusters. Gap statistics denoted as

$Gap_n(k)$ can be calculated using the expression in Equation 2 [23,24];

$$Gap_n(k) = \sum_n\{\log(\delta_k)\} - \log(\delta_k) \quad (1)$$

Where, δ_k represents the degree of clustering which is based on $WCSS$ which can be calculated using the expression in Equation 3 [23,24];

$$WCSS = \sum_{x_i \in C_k} \sum_{y_i \in C_k} \|x_i - y_i\|^2 \quad (2)$$

The procedure to determine optimal number of cluster using gap statistics is presented in Algorithm 2 and it shows that $WCSS$, in this case is calculated by using the $inertia$_ property of $KMeans$ method.

Algorithm 2: Determination of optimal number of clusters based on Gap Statistics method

1: **Begin**
2: Define null reference
3: Compute the cluster congestion
4: Group the reference data set with different number of clusters
5: Compute the congestion average on the dataset
6: Compute Gap statistics based on Equation 1
7: **End**

2.2 DETERMINATION OF OPTIMAL NUMBER OF CLUSTERS USING ELBOW METHOD

The Elbow method presents a simple way to determine the optimum number of clusters for a given set data items to be clustered. In this method, a guess number of clusters, n is initially chosen, the clustering algorithm, which in this study is the K-means method is used to cluster the data items in the n different clusters. The total of the within-cluster sum of square (WSS) is computed for the n number of clusters used in the clustering. The number of clusters, np that results in the minimum value of WSS is deem as the optimal value.

So, the Elbow method starts with initial n (number of clusters, say 2, determine the sum of WSS for the two clusters, increase n by 1, compute WSS again. If the value of WSS obtained with n+1 is the same or approximately no significant difference is observed, then n is taken as the optimal number of clusters required.

If on the other hand, WSS for n+1 is greater than WSS for n, the value of n is incremented, the clustering is performed and WSS is compute. The process is repeated until the minimum value of WSS is obtained and the value of n corresponding to the minimum value of WSS is take as the optimal value.

Mathematically, let WSS_j denote the within-cluster sum of square for the jth cluster where there are n number of clusters. Then, $WSS(n)$ is defined as [25,26];

$$WSS(n) = \sum_{j=1}^{j=n}(WSS_j) \quad (3)$$

The Elbow method will compute $WSS(n)$ for different values of n and will adopt n for which $WSS(n)$ is minimal.

A simple procedure for the optimum number of clusters determination using the Elbow method and K-means algorithm is presented as Algorithm 1.

Algorithm 1.

Step 1: Input nmax // the maximum number of clusters to be considered.

Step 2: Initialize n = 1 // n is the guess optimum number of clusters to be considered at the moment

Step 3: Compute WSSn // WSSn is the sum of within-cluster sum of square (WSS) for n
Step 4: n = n + 1
Step 5: If n ≤ nmax Then Gotp Step 3 Else Goto Step 6
Step 6: Plot the graph of WSSn versus n
Step 7: Locate the knee or sharp bend in the curve of WSSn versus n . The value of n at the knee point is the optimum number of clusters.
Step 8: End

2.3 SIMULATION OF THE GAP STATISTICS AND ELBOW METHODS

A program developed using Python 3 in Pycharm development environment was used to simulate the Gap Statistics and Elbow algorithms. The case study considered has an area, A with dimensions of 800m × 800m which amounts to a 640 km^2 area. A total of 5000 sensor nodes were randomly distributed within the 640 km^2 area as depicted in Figure 1. The simulation program was separately implemented for the Gap Statistics and Elbow methods.

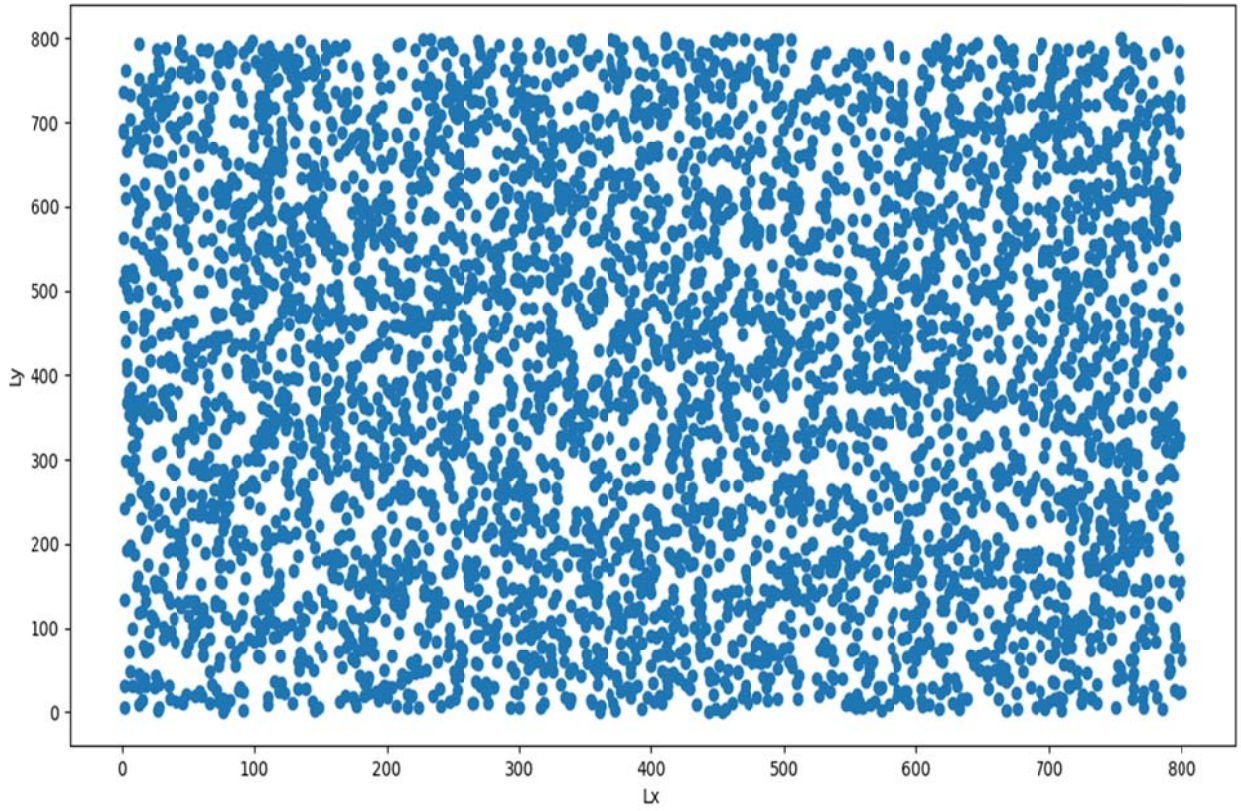

Figure 1: The 5000 sensor nodes randomly distributed within the 640 km^2 area

3 RESULTS AND DISCUSSION

3.1 Evaluation of Number of Cluster Determination Based on Elbow Method

The results obtained for the Elbow method is presented in Figure 2. The results show that the Elbow method gave optimal number of clusters as four; as shown in Figure 2. It means that each of the 5000 sensors nodes are grouped into four clusters given as;$Cluster$ 0, $Cluster$ 1, $Cluster$ 2 and $Cluster$ 3, as shown in Figure 2.

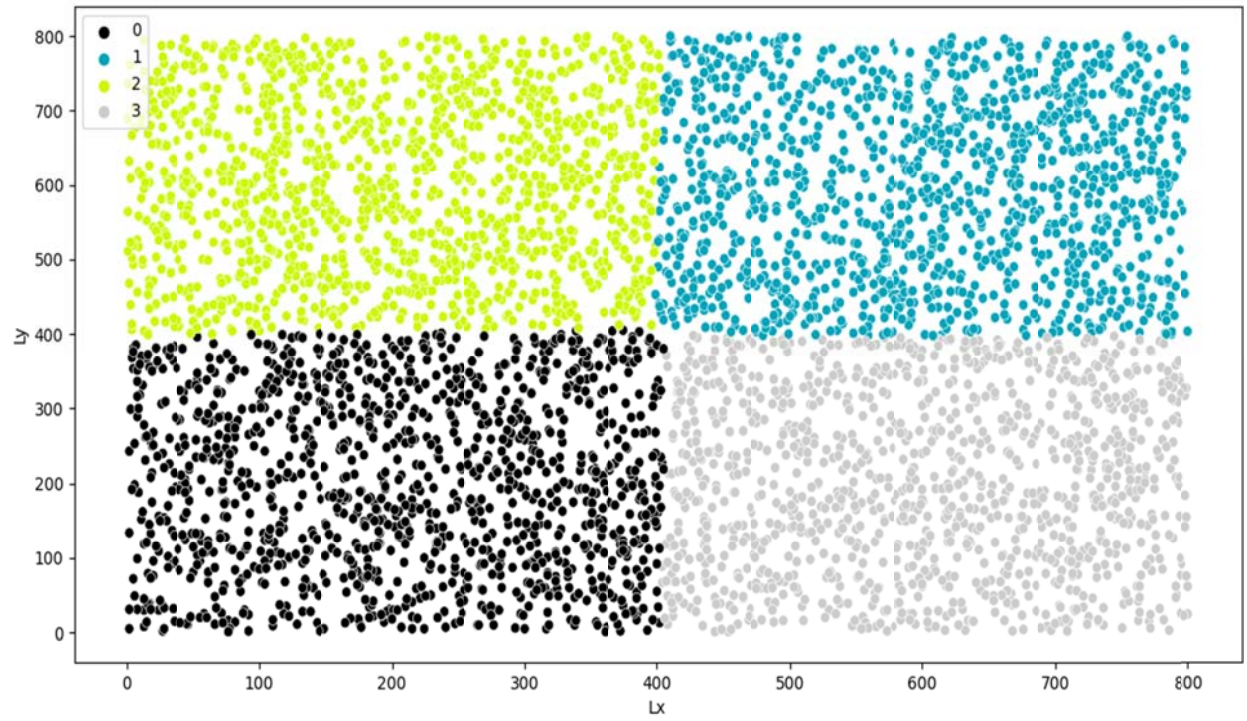

Figure 2: Results of the Elbow method for optimum number of cluster determination

3.2 EVALUATION OF NUMBER OF CLUSTER DETERMINATION BASED ON GAP STATISTICS METHOD

The results obtained for the Gap Statistics method is presented in Figure 3. The results show that the Gap Statistics method gave optimal number of clusters as five; as shown in Figure 3. It means that each of the 5000 sensors nodes are grouped into five clusters given as; $Cluster\ 0$, $Cluster\ 1$, $Cluster\ 2$, $Cluster\ 3$ and $Cluster\ 4$, as shown in Figure 2.

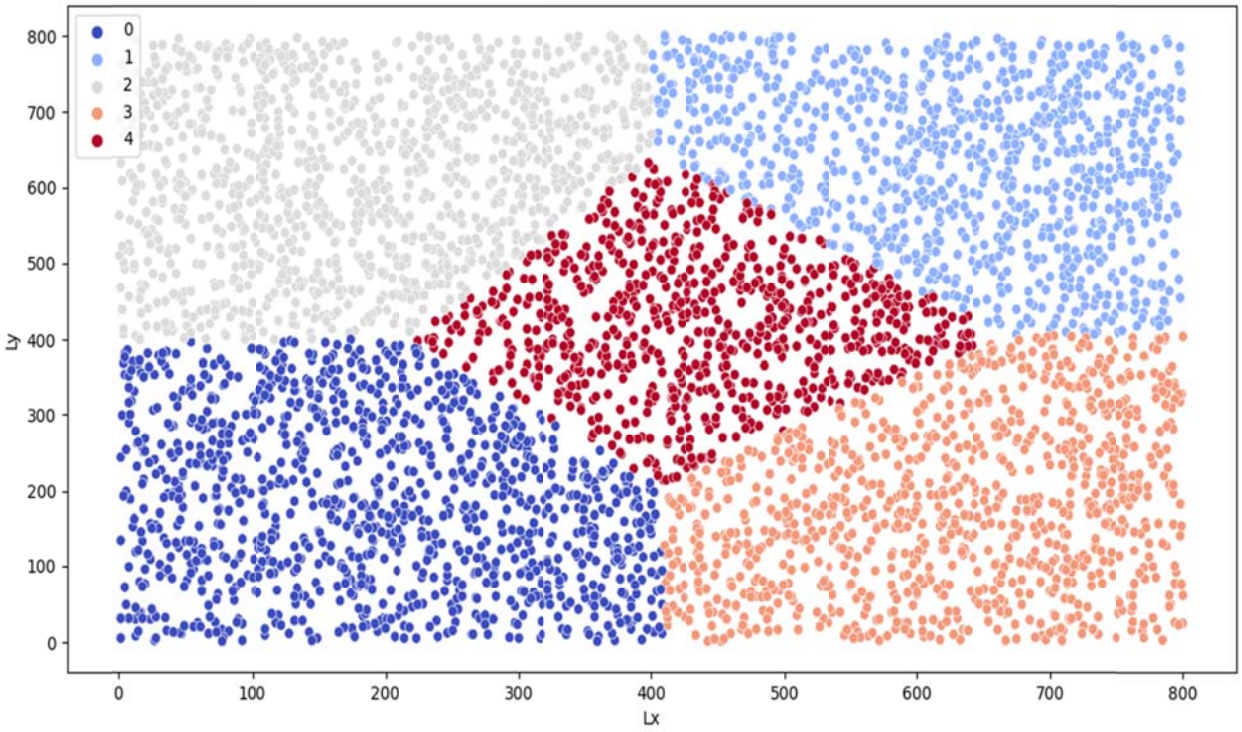

Figure 3: Figure 2: Results of the Gap Statistics method for optimum number of cluster determination

3.3 COMPARISON OF METHOD USED FOR DETERMINATION OF NUMBER OF CLUSTERS

The results presented by Elbow method in Figure 2, the optimal number of clusters was four. However, Gap Statistics method gave five as the optimal number of clusters. The performance of the two methods, the Elbow method and the Gap statistics method execution times are compared as shown in Figure 4. The results show that it took 5.8 seconds for the Elbow method to determine the optimal number of clusters whereas it took the Gap statistics method 5.3 seconds. There is therefore, about 0.5 seconds gained by using the Gap statistics method which is about 8.62069 % improvement in the implementation time over the Elbow method.

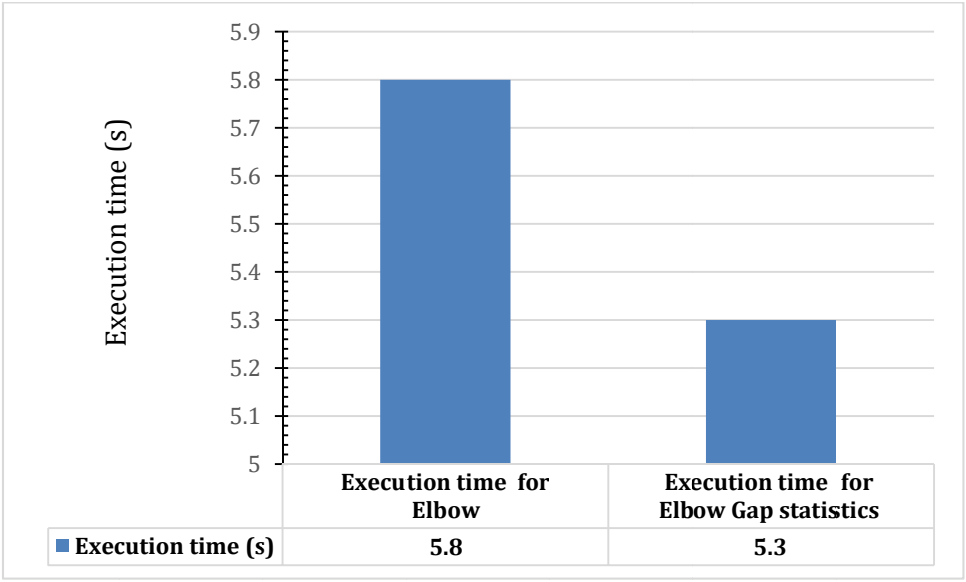

Figure 4:The execution time comparison of the two methods, the Elbow method and the Gap statistics method

4. CONCLUSION

The paper presented two methods for determination of the optimal number of clusters to be used in sensor node clustering. The two methods are the Elbow method and the Gap statistics method. The two methods were simulated using a python program and the results show that the gap statistics gave five as the optimal number of clusters required whereas the Elbow method gave four. Also, the results show that the Gap statistics method execution time is better than that of the Elbow method. Hence, it is better to use the Gap statistic method given its fact implementation time.

REFERENCES

1. Kandris, D., Nakas, C., Vomvas, D., & Koulouras, G. (2020). Applications of wireless sensor networks: an up-to-date survey. *Applied system innovation*, *3*(1), 14.
2. **U. Ukommi**, U. Akpan and S. Garba, "Content-Aware Adaptive Mechanism for Improved Digital Multimedia Delivery over Wireless Channel", Advances in Image and Video Processing Journal, Society for Science and Education, UK, Vol.4, No.2, PP. 31-36, April 2017.
3. **U. Ukommi** and Emmanuel Ubom, "Hybrid Approach for Efficient Distribution of Television Applications over Wireless Channel", International Journal of Innovative Research in Science and Technology, USA, Vol.4, No.2, March 2016.
4. Priyadarshi, R., Gupta, B., & Anurag, A. (2020). Deployment techniques in wireless sensor networks: a survey, classification, challenges, and future research issues. *The Journal of Supercomputing*, *76*, 7333-7373.
5. Ramírez-Moreno, M. A., Keshtkar, S., Padilla-Reyes, D. A., Ramos-López, E., García-Martínez, M., Hernández-Luna, M. C., ... & Lozoya-Santos, J. D. J. (2021). Sensors for sustainable smart cities: A review. *Applied Sciences*, *11*(17), 8198.
6. U. Ukommi, "Smart Broadcast Technique for Improved Video Applications over Constrained Networks", International Journal of Advanced Computer Science and Applications, USA, Vol.4, No.10, 2013. https://thesai.org/Downloads/Volume4No10/Paper_2-Smart_Broadcast_Technique_for_Improved_Video_Applications.pdf
7. Syed, A. S., Sierra-Sosa, D., Kumar, A., & Elmaghraby, A. (2021). IoT in smart cities: A survey of technologies, practices and challenges. *Smart Cities*, *4*(2), 429-475.
8. Ghazal, T. M., Hasan, M. K., Alshurideh, M. T., Alzoubi, H. M., Ahmad, M., Akbar, S. S., ... & Akour, I. A. (2021). IoT for smart cities: Machine learning approaches in smart healthcare—A review. *Future Internet*, *13*(8), 218.
9. Azarhava, H., & Niya, J. M. (2020). Energy efficient resource allocation in wireless energy harvesting sensor networks. *IEEE Wireless Communications Letters*, *9*(7), 1000-1003.
10. Pereira, F., Correia, R., Pinho, P., Lopes, S. I., & Carvalho, N. B. (2020). Challenges in resource-constrained IoT devices: Energy and communication as critical success factors for future IoT deployment. *Sensors*, *20*(22), 6420.
11. Olufemi, O.I., Ukommi, U. (2024). Evaluation of Energy Consumption and Battery Life Span for LoRa IoT Multisensor Node for Precision Farming Application. Signals and Communication Technology. Springer Nature, Switzerland. 153-162. https://link.springer.com/chapter/10.1007/978-3-031-53935-0_15
12. Essien, A., Ukommi, U., & Ubom, E. (2024). Downlink Power Budget and Bit Error Analysis for LoRa-Based Sensor Node-to-Satellite Link in the Industrial, Scientific and Medical Frequency Bands. Signals and Communication Technology. Springer Nature, Switzerland. 143-152. https://link.springer.com/chapter/10.1007/978-3-031-53935-0_14
13. Oduoye, O, Ukommi, U & Ubom, E (2023). Comparative Analysis of Transceiver Payload Size Impact on The Performance of LoRaBased Sensor Node. Science and Technology Publishing (SCI & TECH), 7(8), 1559-1563.
14. Shahraki, A., Taherkordi, A., Haugen, Ø., & Eliassen, F. (2020). Clustering objectives in wireless sensor networks: A survey and research direction analysis. *Computer Networks*, *180*, 107376.
15. Han, Y., Li, G., Xu, R., Su, J., Li, J., & Wen, G. (2020). Clustering the wireless sensor networks: a meta-heuristic approach. *IEEE Access*, *8*, 214551-214564.
16. Alghamdi, T. A. (2020). Energy efficient protocol in wireless sensor network: optimized cluster head selection model. *Telecommunication Systems*, *74*(3), 331-345.
17. Shahraki, A., Taherkordi, A., Haugen, Ø., & Eliassen, F. (2020). Clustering objectives in wireless sensor networks: A survey and research direction analysis. *Computer Networks*, *180*, 107376.

18. Amutha, J., Sharma, S., & Sharma, S. K. (2021). Strategies based on various aspects of clustering in wireless sensor networks using classical, optimization and machine learning techniques: Review, taxonomy, research findings, challenges and future directions. *Computer Science Review*, *40*, 100376.
19. Shahraki, A., Taherkordi, A., Haugen, Ø., & Eliassen, F. (2020). A survey and future directions on clustering: From WSNs to IoT and modern networking paradigms. *IEEE Transactions on Network and Service Management*, *18*(2), 2242-2274.
20. Sagala, N. T., & Gunawan, A. A. S. (2022). Discovering the optimal number of crime cluster using elbow, silhouette, gap statistics, and nbclust methods. *ComTech: Computer, Mathematics and Engineering Applications*, *13*(1), 1-10.
21. Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on Wireless Communications and Networking*, *2021*, 1-16.
22. Onumanyi, A. J., Molokomme, D. N., Isaac, S. J., & Abu-Mahfouz, A. M. (2022). AutoElbow: An automatic elbow detection method for estimating the number of clusters in a dataset. *Applied Sciences*, *12*(15), 7515.
23. Yang, J., Lee, J. Y., Choi, M., & Joo, Y. (2019, December). A new approach to determine the optimal number of clusters based on the gap statistic. In *International Conference on Machine Learning for Networking* (pp. 227-239). Cham: Springer International Publishing.
24. Yang, J., Lee, J. Y., Choi, M., & Joo, Y. (2019, December). A new approach to determine the optimal number of clusters based on the gap statistic. In *International Conference on Machine Learning for Networking* (pp. 227-239). Cham: Springer International Publishing.
25. Umargono, E., Suseno, J. E., & Gunawan, S. V. (2020, October). K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. In *The 2nd international seminar on science and technology (ISSTEC 2019)* (pp. 121-129). Atlantis Press.
26. Nanjundan, S., Sankaran, S., Arjun, C. R., & Anand, G. P. (2019). Identifying the number of clusters for K-Means: A hypersphere density based approach. *arXiv preprint arXiv:1912.00643*.