Statistical Analysis Of The Mean Euclidian Distance Of The Sensor Nodes In A Clustered Network Using Incremental Batch K-Means Approach

Maryam Sufiyanu Masari[1]
Department of Cyber Security
Air Force Institute of Technology, Kaduna, Nigeria
Msufiyanu90@gmail.com

Loretta Onyinye Ilori[2]
Department of Information and Communication Technology
Air Force Institute of Technology, Kaduna, Nigeria
lorettaonyinye@gmail.com

Ikenna Calistus Dialoke[3]
Department of Information and Communication Technology
Air Force Institute of Technology, Kaduna, Nigeria
ikennadialoke@yahoo.com

***Abstract*— In this paper, statistical analysis of mean Euclidian distance of sensor nodes in a clustered network using Incremental Batch K-Means approach is presented. The mean Euclidian distance of clustered sensor nodes is determined using Incremental Batch K-Means (IBKM) approach. Statistical analysis is also conducted for describing the clustered sensor nodes Euclidian distance distribution category and the impact of the statistical parameters on the network implementation. The case study sensor network consists of 2000 sensor nodes randomly located within a network coverage area of 1000 m by 1000 m. The IBKM algorithm was used to group the sensor nodes into 6 clusters. The result of the IBKM algorithm implementation show that the iteration converged to the optimal cluster head (centroid sensor nodes) placement with 6 clusters at about the 22nd iteration. The results also show that cluster 0 has the largest Mean Euclidian distance (MED) value of 1730.8 m while cluster 2 has the lowest MED value of 1328.7 m. The entire sensor network has mean MED of 1566.1 m. In addition, the results show that Euclidian distance of the sensor nodes are normally distributed with mean of 1566.09 m and standard deviation of 651.57 m. The MED has a range of 1698.33 m with Mean Absolute Deviation (MAD) of 535.28 m, Root Mean Square (RMS) of 1696.17 m and Standard Error of Mean of 14.57 m. In all, sensor node cluster 0 is the critical cluster with potential highest energy consumption due to the largest MED value.**

Keywords*— *Clustering Algorithm, Mean Euclidian Distance, Sensor Nodes, Clustered Network, Incremental Batch K-Means

1. INTRODUCTION

Today, sensors and sensor networks are the bedrock of smart applications [1,2,3]. Smart applications use the sensor nodes to monitor the environment or system and through the use of communication mechanism, the sensor relay the captured data to remote systems or server [4,5,6]. In many cases, large number of sensors are deployed and base stations or gateways are needed to connect the sensors to the sensor nodes to the remote server [7,8,9]. In such cases, the placement of each of the base stations in each of the clusters in clustered network requires the use of clustering algorithm. The clustering algorithm implementation will give rise to a given mean distance per cluster which is considered to be optimal [10,11].

In this paper, the mean Euclidian distance (MED) realized in a clustered sensor network using the Incremental Batch K-Means (IBKM) algorithm is studied [12,13,14]. The study focused on statistical analysis of the MED realized when the IBKM is implemented repeatedly of a given set of sensor nodes distributed within a given network area. The study provides analytical evidence of the variations in the optimal MED realized with the IBKM. The study uses statistical approach to determine how the Euclidian distance of the nodes are distributed relative to the MED. It also determine whether the Euclidian distance of the nodes are normally distributed when the IBKM algorithm is repeatedly implemented on the same set of sensor nodes in a given network area.

2. METHODOLOGY

2.1 The algorithm for the Incremental Batch K-Means (IBKM) approach

The mean Euclidian distance of clustered sensor nodes is determined for a sensor network where the sensors are clustered using Incremental Batch K-Means (IBKM) approach. The essence of the study is to know the cluster with the maximum mean Euclidian distance which implies that cluster will expand the maximum energy in the network. Also, the statistical analysis is conducted to determine various statistical parameters for describing the clustered sensor network and the impact of such parameters on the network implementation.

The algorithm for the IBKM is presented in Algorithm 1 [12].

Step 1: Read in the data points $x_i \in X$ or i= 1, 2, 3… n where n is the number of data points

Step 2: Initialize iteration counter, t =0

Step 3: Randomly select one initial centroid, c_0

Step 4: For each k $\in \{2^t - 1, \dots, 2^{t+1} - 2\}$, *split the* cluster c_k into clusters c_{2k+1} and c_{2k+2}

Step 5: Update the centroids $2^t - 1, \dots, 2^{t+1} - 2$ using the classical K-Means (or Batch K-Means) algorithm. In this case, the entire data points, $x_i \in X$ or i= 1, 2, 3… n are clustered.

Step 6: Update the iteration counter, t = t + 1

Step 7: Repeat step 4, step 5 and step 6 until t=$\log_2 K$

2.2 The analytical expressions for the statistical parameters related to the Euclidian distance of the sensor nodes from the cluster head location

The following parameters are used to describe the parameters used in computing the mean Euclidian distance (MED);

xC_k is the cluster k cluster head location x coordinate where k $\in \{1,2,3, \dots Kn\}$, Kn represented the number of clusters

yC_k is the cluster k cluster head location y coordinate where k $\in \{1,2,3, \dots Kn\}$

$x_{j,k}$ is the x coordinate of the sensor node in cluster k where j $\in \{1,2,3, \dots Jn\}$, Jn represented the number of sensor nodes clustered in cluster k

$y_{j,k}$ is the y coordinate of the sensor node in cluster k where j $\in \{1,2,3, \dots Jn\}$

$d_{j,k}$ is Euclidian distance for each sensor node in cluster k

dm_k is the MED for cluster k

dm_{All} is the MED for all the clusters

Then,

$$d_{AVG} = \left(\frac{1}{Kn}\right)\left(\sum_{i=1}^{i=Kn}(dm_k)\right) \quad (1)$$

Where,

$$dm_k = \left(\frac{1}{Jn}\right)\left(\sum_{i=1}^{i=Jn}(d_{j,k})\right) \quad (2)$$

$$d_{j,k} = \sqrt{\left(xC_k - x_{j,k}\right)^2 + \left(yC_k - y_{j,k}\right)^2} \quad (3)$$

The standard deviation of $d_{j,k}$ for all k is denoted as S where n is the number of sensor nodes in the network;

$$S = \frac{\sqrt{\sum_{k=1}^{k=Kn}\left\{\sum_{j=1}^{j=Jn}\left(d_{j,k} - dm_k\right)^2\right\}}}{n-1} \quad (4)$$

The range, R is defined as ;

$$R = maximum\left(d_{j,k}\right) - minimum\left(d_{j,k}\right) \quad (5)$$

The Mean Absolute Deviation (MAD) is defined as;

$$\text{MAD} = \frac{\sqrt{\sum_{k=1}^{k=Kn}\left\{\sum_{j=1}^{j=Jn}\left(d_{j,k} - dm_k\right)\right\}}}{n} \quad (6)$$

The Root Mean Square (RMS) is defined as;

$$RMS = \frac{\sqrt{\sum_{k=1}^{k=Kn}\left\{\sum_{j=1}^{j=Jn}\left(d_{j,k}\right)^2\right\}}}{n} \quad (7)$$

The Standard Error of the Mean ($SE_{\bar{x}}$) is defined as;

$$RMS = \frac{S}{\sqrt{n}} \quad (8)$$

The percentile, $P_{j,k}$ (%) for the $d_{j,k}$ is computed as follows;

$$P_{j,k}\ (\%) = \frac{(d_{j,k})(100)}{maximum(d_{j,k})} \quad \text{for all k} \in \{1,2,3, \dots Kn\} \text{ and j } \in \{1,2,3, \dots Jn\}, \quad (9)$$

The normal distribution of the $P_{j,k}$ is determined and plotted in Microsoft Excel.

2.3 The Case Study Sensor Network Dataset

The case study sensor network consists of 2000 sensor nodes randomly located within a network coverage area of 1000 m by 1000 m as depicted in Figure 1. The IBKM algorithm was used to group the sensor nodes into 6 clusters. The Euclidian distance parameters were computed along with the various statistical parameters outlined in Equation 1 to Equation 9.

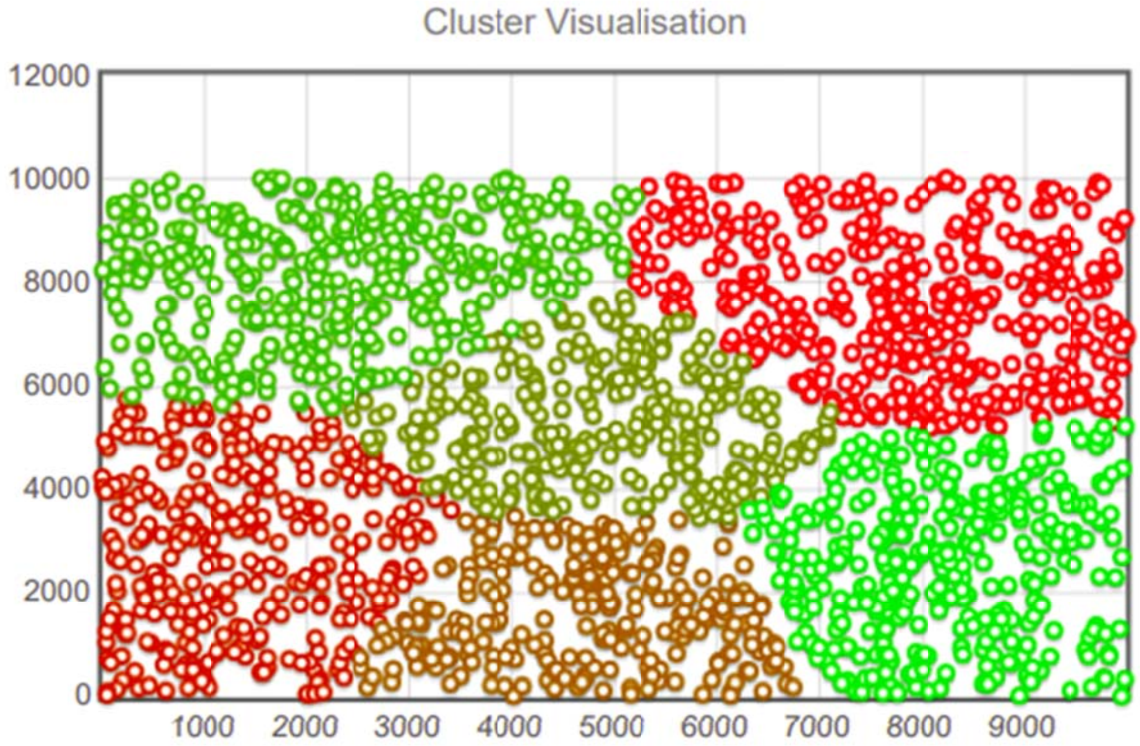

Figure 1 The case study sensor network consisting of 2000 sensor nodes randomly located within a network coverage area of 1000 m by 1000 m

3. RESULTS AND DISCUSSIONS

The implementation of the IBKM algorithm gave the graph of the changes in the sensor node cluster location for the various iterations of the as IBKM algorithm as presented in Figure 2. The graph show that the iteration converged to the optimal cluster head (sensor nodes) placement with 6 clusters at about the 22nd iteration. The cluster head sensor node location X and Y coordinates, $(xC_k, yC_k\text{n})$ for the 6 clusters in the sensor network are shown in Table 1. The results of the mean value, maximum and minimum values of the Euclidian distance for the sensor nodes in each of the 6 clusters are presented in Table 2.

The results in Table 2 show that cluster 0 has the largest Mean Euclidian distance (MED) value of 1730.8 m while cluster 2 has the lowest MED value of 1328.7 m. The entire sensor network has mean MED of 1566.1 m. The plot of the results on the percentile for the Euclidian distance, $(j,)$ versus the number of sensor node that has their Euclidian distance in the range with respect to the maximum Euclidian distance in the entire sensor network are presented in Figure 2. The normal distribution graph plot of the percentile is shown in Figure 3. The graph in Figure 3 shows that the Euclidian distance of the sensor nodes are normally distributed with mean (d_{AVG}) of 1566.09 m and standard deviation of 651.57 m. The range is 1698.33 m with Mean Absolute Deviation (MAD) of 535.28 m, Root Mean Square (RMS) of 1696.17 m and Standard Error of Mean of 14.57 m. In all, sensor node cluster 0 is the critical cluster with potential highest energy consumption due to the largest MED value.

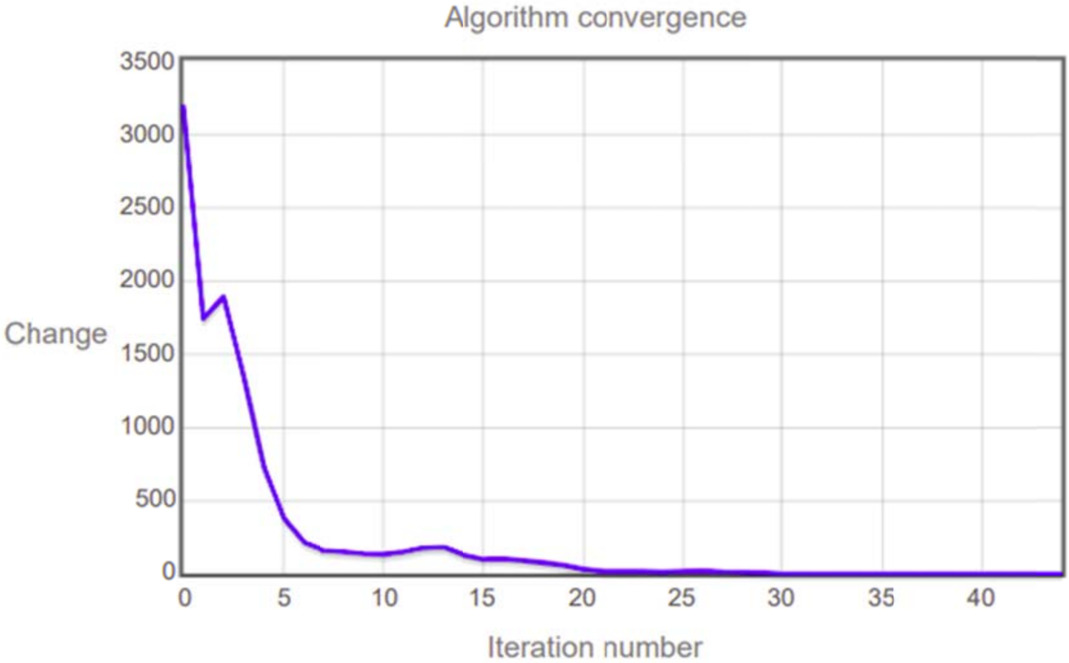

Figure 2 The graph of the changes in the sensor node cluster location for the various iterations of the as IBKM algorithm

Table 1 The cluster head sensor node location X and Y coordinates, $(xC_k, yC_k\text{n})$ for the 6 clusters in the sensor network

Cluster Number	Cluster head sensor node location X coordinate, xC_kn (m)	Cluster head sensor node location y coordinate, yC_kn (m)
0	7805.297	7602.153
1	1395.367	2860.309
2	4728.416	1619.031
3	4911.596	5301.279
4	2313.598	8165.768
5	8208.201	2585.192

Table 2 The mean value, maximum and minimum values of the Euclidian distance for the sensor nodes in each of the 6 clusters

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Min	Max	Average
No. of nodes	411	324	257	287	383	338	257	411	333.3
Percentage of total nodes	20.6	16.2	12.9	14.4	19.2	16.9	12.9	20.6	20.0
Maximum Distance	109.7	236.8	141.6	173.6	128.0	66.3	109.7	236.8	171.2
Minimum Distance	3330.4	3115.2	2565.3	2497.8	3255.0	3178.5	2497.8	3330.4	3588.4
Average Distance	1730.8	1649.3	1328.7	1424.0	1598.1	1551.0	1328.7	1730.8	1566.1

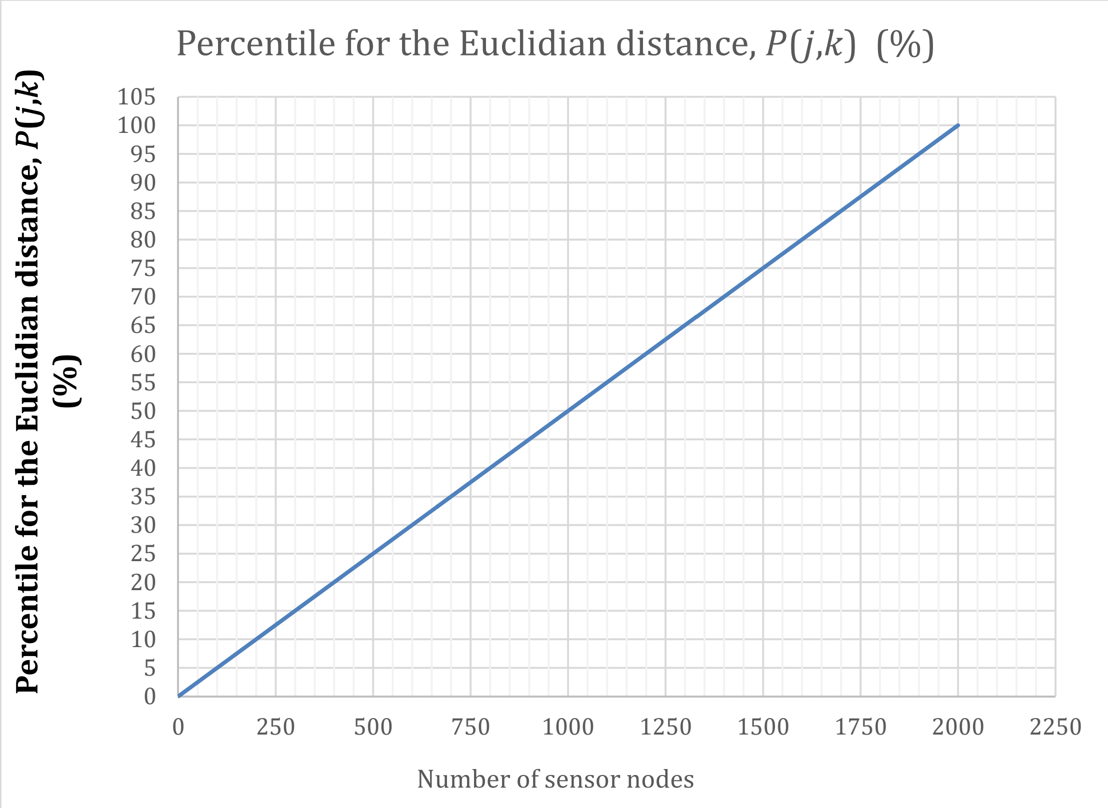

Figure 2 The percentile for the Euclidian distance, $(j,)$ versus the number of sensor node that has their Euclidian distance in the range with respect to the maximum Euclidian distance in the entire sensor network

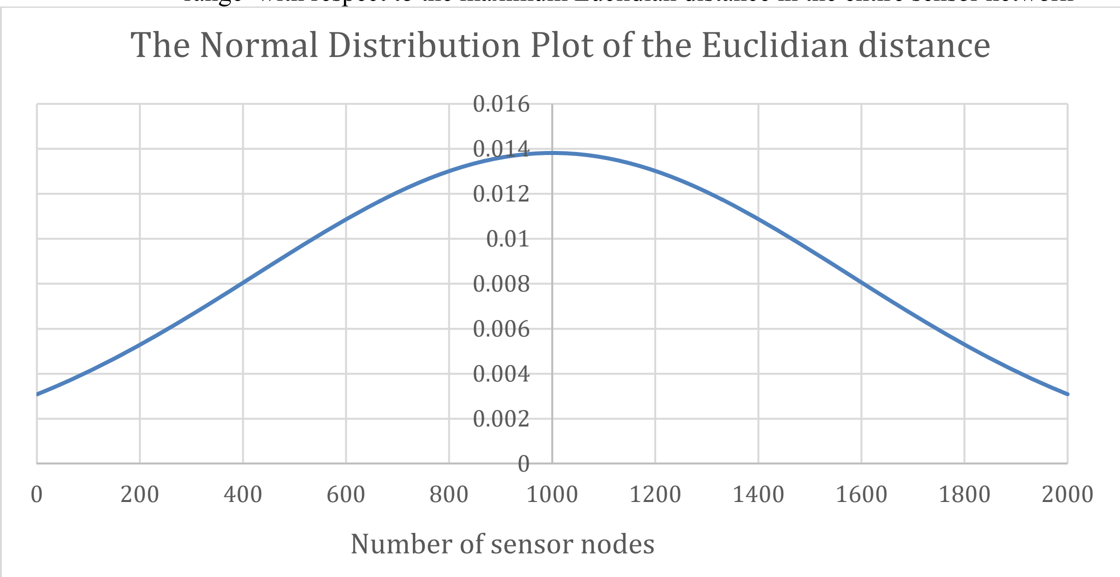

Figure 3 The normal distribution graph plot of the percentile

4. CONCLUSION

The clustering of sensor node using Incremental Batch K-Means (IBKM) approach is studied. The statistical distribution of the Euclidian distance of the sensor nodes within each cluster is studied to identify the critical cluster in the network. Some key statistical parameters of the distribution are also considered. In all, it was observed from the results that out of the 6 clusters generated from the IBKM algorithm, cluster 0 has the highest MED value and hence is the critical cluster when considering the energy consumption and battery lifespan for the sensor nodes.

REFERENCES

[1] Hart, J. K., & Martinez, K. (2020). Sensor networks and geohazards. *Environ. Earth Sci.*

[2] Gamperl, M., Singer, J., & Thuro, K. (2021). Internet of things geosensor network for cost-effective landslide early warning systems. *Sensors*, *21*(8), 2609.

[3] Li, C. C., Ramanna, V. K., Webber, D., Hunter, C., Hack, T., & Dezfouli, B. (2021). Sensifi: A wireless sensing system for ultrahigh-rate applications. *IEEE Internet of Things Journal*, *9*(3), 2025-2043.

[4] Talal, M., Zaidan, A. A., Zaidan, B. B., Albahri, A. S., Alamoodi, A. H., Albahri, O. S., ... & Mohammed, K. I. (2019). Smart home-based IoT

for real-time and secure remote health monitoring of triage and priority system using body sensors: Multi-driven systematic review. *Journal of medical systems*, *43*, 1-34.

[5] Abdelhafidh, M., Fourati, M., Fourati, L. C., & Chouaya, A. (2019). Wireless sensor network monitoring system: architecture, applications and future directions. *International Journal of Communication Networks and Distributed Systems*, *23*(4), 413-451.

[6] Majid, M., Habib, S., Javed, A. R., Rizwan, M., Srivastava, G., Gadekallu, T. R., & Lin, J. C. W. (2022). Applications of wireless sensor networks and internet of things frameworks in the industry revolution 4.0: A systematic literature review. *Sensors*, *22*(6), 2087.

[7] Radovici, A., Culic, I., Rosner, D., & Oprea, F. (2020). A model for the remote deployment, update, and safe recovery for commercial sensor-based iot systems. *Sensors*, *20*(16), 4393.

[8] Bravo-Arrabal, J., Fernández-Lozano, J. J., Serón, J., Gomez-Ruiz, J. A., & García-Cerezo, A. (2021). Development and implementation of a hybrid wireless sensor network of low power and long range for urban environments. *Sensors*, *21*(2), 567.

[9] Nourildean, S. W., Hassib, M. D., & Mohammed, Y. A. (2022). Internet of things based wireless sensor network: a review. *Indones. J. Electr. Eng. Comput. Sci*, *27*(1), 246-261.

[10] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1), e0210236.

[11] Aldino, A. A., Darwis, D., Prastowo, A. T., & Sujana, C. (2021). Implementation of K-means algorithm for clustering corn planting feasibility area in south lampung regency. In *Journal of Physics: Conference Series* (Vol. 1751, No. 1, p. 012038). IOP Publishing.

[12] Abernathy, A., & Celebi, M. E. (2022). The incremental online k-means clustering algorithm and its application to color quantization. *Expert Systems with Applications*, *207*, 117927.

[13] Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2004). An incremental K-means algorithm. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *218*(7), 783-795.

[14] Chakraborty, S., & Nagwani, N. K. (2014). Performance evaluation of incremental K-means clustering algorithm. *arXiv preprint arXiv:1406.4737*.