iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.57868

Performance of ML-Based Unsupervised Clustering Algorithms for WSN Node Clustering

M. Manoranjani[1], Dr. S. Sukumaran[2]
[1]*Ph. D Research Scholar*, [2]*Associate Professor*
[1, 2]*Erode Arts and Science College, Erode, Tamilnadu*

Abstract: Wireless Sensor Networks are generally deployed in dynamically changing environment. When compared to common wired network nodes, WSN nodes must do more work. Since WSN devices are battery powered, so power management is a challenge. Clustering is one solution that has been proposed to alleviate the issue of limited power. Clustering is the most important method stabilizes the lifetime of the network. It entails the aggregation of sensor nodes into clusters and cluster head is picked out from all the clusters. Clustering is implemented in wireless sensor networks through the Machine Learning techniques. In wireless sensor networks, machine learning algorithms have an important role in cluster head formation and maintain the stability of the nodes in the cluster. Machine Learning approaches used in wireless sensor networks can be classified as Supervised learning, Unsupervised learning, and Reinforcement learning. Among these learning techniques, unsupervised learning deals with different clustering algorithms such as k-means, K-medoids, Fuzzy C-means, hierarichal-based, and SOM. This paper evaluates the performance of the variants of k-Means (kM) and Fuzzy C-means (FCM) algorithms in terms of the clustering and accuracy. This paper imparts performance analysis of different clustering algorithms in machine learning applied for wireless sensor networks. From the analysis, the Fuzzy C-Means algorithm found to be more suitable for node clustering in WSN.

Keywords: WSN, Clustering, Machine Learning techniques, k-Means, FCM.

I. INTRODUCTION

WSNs monitor dynamic environments that change rapidly over time. During data distribution each node communicates with Base station through a single hop or multi hop data transfer. As the continuous process of data transmission, the nodes having the more significant distance consume their resources quickly than other nodes, and hence to solve this issue, the clustering method is used in WSNs with several nodes [2]. Clustering and routing are the two mechanisms utilized in WSNs to extend network lifetime [3]. Clustering involves the formation of groups of sensor nodes known as clusters, and the selection of a Cluster Head (CH) inside each cluster, who is a highly qualified node. The CH collects data from the cluster members and transmits it to the Base Station (BS). To BS, data transmission can be either single-hop or multi-hop. When data packets go from source to destination via single-hop transmission, only one networking device is present. Data packets in multi-hop transmission go from source to destination via more than one networking device. Multi-hop transmission is most typically employed in large-scale networks. Clustering nodes in WSNs improves scalability, reduces routing time, and increases energy efficiency. However, in wireless sensor networks, external influences and dynamic changes effect cluster head selection, routing, latency, localization, QoS, fault detection, dependability, and security. As a result of this repercussion, the network does not function well in a dynamic and complex environment.

To overcome this problem, Machine Learning algorithms are used. Machine Learning is a branch of artificial intelligence which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. There are four basic types of machine learning: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The power of machine learning rests in its capacity to give universal solutions via an architecture that can learn to improve its performance. Because of its interdisciplinary nature, it is important in many fields, including engineering, medicine, and computers. Recent improvements in machine learning have been used to overcome a variety of problems in WSNs. Using ML not only increases the performance of WSNs, but it also reduces the need for human intervention or re-programming. Some of the performance metrics are used to evaluate the machine learning techniques and obtain which algorithm is reliable for making clustering in WSN. Real-world data is messy so Data Preprocessing is an important step in Machine Learning to transforms the data in a format that can be understood and analyzed with less computation complexity [4]. The perfect preprocessed data is even more important than the most powerful algorithm. In this paper, Section II describes the related works of unsupervised machine learning algorithms, Section III shows the experimental results of clustering algorithms and Section IV describe the conclusion about the clustering problems in WSN and suggestion for future work.

II. RELATED RESEARCH WORKS

A substantial portion of sensor network research has concentrated on energy-efficient clustering-based routing algorithms [9]. We looked at a range of tactics in this study, highlighting a few of them. The sensor nodes in the network can be configured to operate as CH either centrally or distributedly. The first employs a BS to handle CH selection, but the second is totally self-organized.

Machine learning is increasingly being utilized to divide the network into clusters, from which CHs are selected depending on predefined criteria. This can be accomplished by employing algorithms like k-means [8,9] and fuzzy c-means [10], which are increasingly being employed in WSNs, IoT, and crowd-sensing applications. To deal with uncertainty in WSNs [11], fuzzy logic-based clustering algorithms were applied. The authors [10] proposed fuzzy logic-based data processing and grouping for WSNs. This technique considers the energy level, bandwidth, and connection efficiency of each node. The proposed work intends to improve network performance in terms of network lifetime, number of live nodes, CH selection time, throughput, and energy utilization.

Amir et.al [3], developed the HCQA approach employs a criterion for determining cluster quality, which can increase inter-cluster and distances while also lowering error rates during clustering. The best cluster head (CH) is chosen using fuzzy logic and numerous criteria such as residual energy, minimum and maximum distances between nodes in each cluster and the base station. The authors [11] proposed Power-Efficient Cluster-based Routing (PECR) is a revolutionary technique that employs K-means clustering, optimal route selection, communication based on energy use, cluster head and primary cluster head change for energy efficiency and longer network longevity. There have been numerous studies on WSN routing protocols with energy efficiency, security, and cluster-based routing. [9-13].

The K-means technique, created by J.B.Mac Queen in 1967, is offered as one of the simplest non-supervised learning algorithms for clustering issues. In [12], a variation of the LEACH algorithm improves the clustering procedure. LEACH's random clustering will be replaced with the k-means clustering technique in the adaptation. This adaptation has improved clustering allocation and cluster features, as well as generated energy efficient clustering to extend the life of WSNs. When the CHs are re-elected, the usage of the k-means method as a clustering strategy ensures flawless grouping and reduces overheads.

A fuzzy-logic-based routing strategy is proposed in [10] for achieving energy-optimized, multi-parameter, and fuzzy routing decisions. On-demand clustering significantly reduces the amount of unneeded control message delivery. Forming clusters of nodes using a modified k-means algorithm [13] with starting centroids based on the geographic region of the network is one method for cluster generation and selection of stable cluster heads. Following that, the cluster heads are chosen using the weighted multi-criterion acceptability formula.

III. MACHINE LEARNING CLUSTERING ALGORITHMS FOR WSN

In this section an introduction to the few of clustering based machine learning algorithms that were used in WSN is provided.

The Machine learning techniques which automatically learns task using example data without being specifically programmed is a class of Artificial Intelligence [14]. In recent enlargements, Machine Learning (ML) techniques have been used to solve different problems in WSN to ensure that good decisions can be done in complex situations. The algorithms in ML generate cost effective approach compared to numerical methods.

A. Unsupervised Learning

Unsupervised learning is a branch of machine learning and artificial neural networks. Unsupervised learning uses unlabeled training data. The purpose of an unsupervised learning method is to assess data density in order to detect commonalities between objects and statistically arrange them [8]. Unsupervised learning methods are commonly employed to handle clustering, dimensionality reduction, and outlier detection problems. The goal of clustering is to group feature vectors based on their qualities. Unsupervised learning makes significant contributions to WSN by addressing difficulties like as connection, anomaly detection, routing, and data aggregation. Unsupervised learning is classified into clustering (k-means, hierarchical, k-medoids, fuzzy-c means, and SOM) and dimensionality reduction (PCA, ICA, and SVD).

1) k-Means Clustering Algorithm

k-Means is the simplest algorithm used for unsupervised clustering. This algorithm partitions the data set into k-clusters using Euclidean distance mean, resulting in maximizing intra-cluster similarity and minimizing inter-cluster similarity. K-means is iterative in nature [12]. It follows the following steps:

a) Generate k points (cluster centers) at random, where k is the desired number of clusters.

b) Determine the distance between each data point and each center and allocate each data point to the center that is closest to it.

c) Determine the new cluster center by arithmetic mean of all data points in the relevant cluster.
d) Repeat step (2) with the new centers. If the data points are assigned, repeat step (3); otherwise, halt the operation.

The distance between the data points is calculated using Euclidean distance defined by

$$\text{Dist } X_1, X_2 = \sqrt{\sum_{i=1}^{n} x_{1i} - x_{2i^2}} \quad (1)$$

k-means clustering is the simplest clustering and useful in WSNs to find optimal cluster heads (CHs) for routing the data towards to the base station. This approach is useful to find the efficient rendezvous points for mobile sink.

2) Fuzzy C-Means Clustering

Bezdek created Fuzzy-C-Means (FCM) clustering, also known as soft clustering, in 1981 using fuzzy set theory, which allocates observations to one or more clusters. It is an unsupervised clustering approach that allows us to construct a fuzzy division from data. The approach is based on a parameter m that represents the degree of fuzziness of the solution. When m is large, the classes get muddled, and all items tend to belong to all clusters. The optimization problem's solutions are determined by the parameter m. That is, different values of m result in distinct partitions.

The Fuzzy C-Means (FCM) algorithm has been the most extensively employed in the clustering process among the fuzzy clustering approaches [9]. FCM aims to reduce the sum of distances between instances and cluster centroids [10]. The goal of WSNs is to group N sensor nodes into k distinct clusters. The FCM objective function for clustering in WSNs can be expressed as follows:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \mu^m \, d(x_i, x_c)^2, i = 1,2, \ldots n, \; j = 1,2, \ldots, k \quad (2)$$

$$\mu_{ij} = \frac{1}{\sum_{j=1}^{k} \left(\frac{d(x_i, c_j)}{d\,(x_i, c_k)}\right)^{\frac{2}{m-1}}} \qquad \mu_{ij} \in [0,1] \quad (3)$$

$$C_j = \frac{\sum_1^n (\mu_{ij})^m \, x_i}{\sum_1^n (\mu_{ij})^m} \quad (4)$$

Where µ is the membership of node i to cluster j, m is the value of fuzzifier is usually chosen as 2 in the most of the application [7], C_j refers to cluster centriod. This function differs from KM with use of weighted squared errors instead of using squared errors only.

3) K-medoids

The K-medoids algorithm is a clustering algorithm related to k-means algorithm. Both algorithms are partitional and both attempts to minimize squared error, the distance between points labeled to be in cluster and a point designated as the center of that cluster. In contrast to the K-means algorithm K-medoids [2] choose data point as centers. K-medoid is a traditional clustering partitioning approach that divides a data set of n objects into k known a priori clusters. The silhouette is an excellent tool for calculating k. It is more resistant to fluctuations and outliers than k -means.

A medoid is an item in a cluster having the least amount of dissimilarity to the other objects in the cluster, i.e. it is the most centrally situated point in the supplied data set. The Partitioning Around Medoids (PAM) algorithm is the most frequent implementation of k -medoid clustering and is as follows:

a) Begin by randomly selecting k of the n data points as the medoids.
b) Match each data point to the nearest medoid.("closest" is defined in this context as any valid distance metric, most often Euclidean distance, Manhattan distance, or Minkowski distance).
c) Swap m and o and compute the total cost of the configuration for each medoid m and each non-medoid data point o.
d) Choose the configuration that has the lowest cost.
e) Repeat steps 2–5 until the medoid shows no change.

The K-medoids algorithm seeks to reduce the sum of the distances between data points and associated cluster medoids[5].

In these algorithms k-means, FCM, k-medoids iterates on the clusters to find the optimal cluster centers. For overlapped datasets, FCM produces better grouping than k-means. It, as well, requires prior knowledge of the number of clusters, as k-means clustering. The FCM has a larger time complexity than the other clustering algorithms, and it is mostly determined by the number of clusters, dimensions, data points, and iterations. This clustering method is utilized in many domains, including pattern recognition, picture segmentation, bioinformatics, and business intelligence. FCM technique used to solve several issues in WSNs such as localization, connectivity, and mobile sink. The FCM has been proved to enhance the network performance in terms of energy consumption

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

IV. RESULTS AND DISCUSSIONS

The variations of k-means and FCM is been compared using various parameters. To evaluate the performance for these algorithms in the formation of a balanced size of clusters, three metrics are used.

A. *Standard Deviation of Mean Square Error STD (MSE) for Intra-Distances [15]*

It calculates the difference in homogeneity for the average intra-distance of each cluster. This measure demonstrates how the average intra-distances of nodes to the cluster's centroid vary from cluster to cluster. It is preferable when the factor is modest, indicating that the intra-distances for clusters are uniform.

$$STD\ (MSE) = \sqrt[2]{\sum |MSE(j) - \mu|^2} \quad J = 1,2,..k \tag{5}$$

Where $STD\ (MSE)$ is mean the standard deviation of Mean Square Error, k is the number of clusters, and μ is the Average of Mean Square Error for distances.

$$MSE\ (j) = \left(\frac{1}{n}\right) * \sum_{i=1}^{n} D\ (x_i, x_c)^2, i = 1,2,..n \qquad c = 1,2,..k \tag{6}$$

The acronym MSE refers to the average of square intra-distances of nodes to the cluster's centroid, n is number of nodes in each cluster, and $D\ (x_i\ , x_c)^2$ square intra distances for node (x_i) to its cluster centroid (x_c) in the cluster (c).

$$\mu = \frac{\sum_{j=1}^{k} MSE}{k} \tag{7}$$

B. *Variation for Clusters Size (V)*

It assesses the dissimilarity of the density of nodes in clusters (number of member nodes in each cluster), with the less the factor indicating a better cluster size. That is, the size of the clusters is balanced.

$$V = \frac{\sum |S_j - \mu|^2}{k} \tag{8}$$

$$\mu = \frac{\sum_{j=1}^{k} S_j}{k} \tag{9}$$

Where Sj refers to cluster size (j) and μ refer to the mean of clusters size[15].

C. *Clusters Size Range (CSR)*

It calculates the ratio of minimum to maximum cluster size. As a result, the size range of clusters should be kept within this range (CSR to 1), with the narrower the range being preferable (near to 1). That suggests there is no significant difference in size between the minimum and maximum cluster sizes.

$$CSR = min\left(\frac{CS_j}{max.\,CS}\right) \tag{10}$$

Table 1 Comparison of Unsupervised Clustering Algorithms in ML for WSN

Clustering Algorithms	Performance Parameters					
	Cluster Objective	Cluster Stability	Cluster Formation	CH selection	Energy Efficiency	MSE
k-Means [1]	Energy saving	Moderate	Random/ Heterogeneous	Min. Distance	Moderate	2.96%
Fuzzy C-Means [9]	Scalability and Connectivity	Moderate	Unique/ Homogeneous	Residual energy	Very low	1.14%
K- medoids	Scalability & Load Balancing	Low	Random/ Homogenous	Threshold function	Moderate	2.38%

In this study, the parameters are used in combination, where it is not adequate to assume that this network has a balanced size of clusters more than other networks based solely on the density of the distribution of the nodes among the clusters, regardless of the homogeneity to the average intra-distances in the clusters, and vice versa. Furthermore, it is critical to establish the size range of clusters, where the volumetric width reflects the difference in size between the network's largest and smallest clusters. The change in cluster sizes is from CSR value to 1, with the narrower (closer to 1) the better. So, these three characteristics are utilized to evaluate the performance of these algorithms in terms of which of these algorithms can build more balanced clusters than another with the random distribution method for nodes in the WSN monitoring area. Matlab is used for simulation, and it is based on the most common cases in the literature, where the number of nodes is 100, the monitoring area is 100*100, and the number of clusters is 5 with 3 iterations. Table 1 also presents a comparison assessment of the various clustered routing Machine Learning algorithms in WSNs based on a few key parameters.

To summarize, recent clustering algorithms have been compared with classical algorithms with different parameters. From the above table, the comparison is among the variants of k-means and Fuzzy C Means clustering algorithms. Different parameters have been chosen such as cluster objective, cluster formation, CH selection, energy efficiency, communication between CH and BS. The complexity of the algorithms also been discussed. The result of comparative study shows that that FCM is much more efficient than k-means. Also, it reveals that the fuzzifier value m = 2 in FCM, which has been widely adopted in many applications, is not a good choice, particularly for sensor nodes with great variation in cluster sizes. Therefore, for data sets with significant uneven distributions in cluster sizes, a smaller fuzzifier value is preferred for FCM clustering, and k-means clustering is a better choice compared with FCM clustering.

V. CONCLUSION

Clustering approach of routing has an advantage of increasing the life time of network as compared to other routing protocols. In this paper, presented some cluster based routing algorithms and their comparative study. The comparative study helps us to analyze that which protocol can be used for which application and which scenario. The results demonstrate that FCM has stronger uniform effect than *k*-means clustering to formation a balanced cluster's size based on these parameters with the random distribution manner for sensor nodes in the monitoring area. Although FCM is superior to KM, but still suffer from the effect of the random nodes deployment condition, where sometimes form imbalanced clusters. This limitation requires proposing assist mechanism to overcome this problem; this will be addressed in future work. At the conclusion and based on the discussion, FCM is a better choice to form a balanced cluster especially when the number of nodes distributed is high along with the big distance of the monitoring area in random nodes deployment in the monitoring area for WSNs.

REFERENCES

[1] Mervat Mustafa Raouf, “Clustering in Wireless Sensor Networks (WSNs)”, Journal Of Baghdad University College Of Economic Sciences- Issue-57- April 2019
[2] J. Amutha, Sandeep Sharma, Sanjay Kumar Sharma, “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”, May 2021, Computer Science Review 40(4):100376, DOI:10.1016/j.cosrev.2021.100376.
[3] Amir AbbasBaradarana, KeivanNavib, “HQCA-WSN: High-quality clustering algorithm and optimal cluster head selection using fuzzy logic in wireless sensor networks”, Fuzzy Set Syst, Nov 2019, DOI: https://doi.org/10.1016/j.fss.2019.11.015
[4] S.Sivasankari, Dr.S.Sukumaran, Dr.S.Muthumarilakshmi “Deep Learning based Weight Guided Wrapper Feature Subset Method for Multiclass Data Classification”, NeuroQuantology |December 2022 | Volume 20 | Issue 16 |Page 5675-5686| doi:10.48047/NQ.2022.20.16.NQ880579
[5] Ahmed Mahdi Jubair, Rosilah Hassan, Azana Hafizah Mohd Aman, “Optimization of Clustering in Wireless Sensor Networks: Techniques and Protocols”, Applied. Science, 2021, 11(23), https://doi.org/10.3390/app112311448.
[6] Kim, Taeyoung; Vecchietti, Luiz Felipe; Choi, Kyujin; Lee, Sangkeum; Har, Dongsoo (2020). Machine Learning for Advanced Wireless Sensor Networks: A Review. IEEE Sensors Journal. doi:10.1109/JSEN.2020.3035846
[7] Padmalaya Nayak, G.K. Swetha , Surbhi Gupta, K. Madhavi, “Routing in wireless sensor networks using machine learning techniques: Challenges and opportunities”, Measurement 178 (2021) 108974, https://doi.org/10.1016/j.measurement.2021.108974
[8] Taeyoung Kim, Luiz Felipe Vecchietti, Kyujin Choi, Sangkeum Lee, “Machine Learning for Advanced Wireless Sensor Networks: A Review”, IEEE Sensors Journal, 1530-437X (c) 2020, DOI 10.1109/JSEN.2020.3035846
[9] Abdulzahra, Ali Mohammed Kadhim and Al-Qurabat, Ali Kadhum M. "A Clustering Approach Based on Fuzzy CMeans in Wireless Sensor Networks for IoT Applications," Karbala International Journal of Modern Science: Vol. 8: Iss. 4, Article 2 (2022). DOI: https://doi.org/10.33640/2405-609X.3259
[10] A. Ali, A. Ali, F. Masud et al., Enhanced Fuzzy Logic Zone Stable Election Protocol for Cluster Head Election (E-FLZSEPFCH) and Multipath Routing in wireless sensor networks, Ain Shams Engineering Journal, 2090-4479 2023, https://doi.org/10.1016/j.asej.2023.102356
[11] Firdous, S.; Bibi, N.; Wahid, M.; Alhazmi, S. Efficient Clustering Based Routing for Energy Management in Wireless Sensor Network-Assisted Internet of Things. Electronics 2022, 11, 3922.

[12] Bouakkaz Fatima, Ali Wided, Guemmadi Sabrina and Derdour Makhlouf, "K-Means Efficient Energy Routing Protocol for Maximizing Vitality of WSNs", Computational Optimization Techniques and Applications, 2021, DOI: 10.5772/intechopen.96567

[13] Amit Gupta, Mahesh Motwani and J. L. Rana, Improved Performance Clustering Using Modified K-Means Algorithm in Mobile Adhoc Networks, International Journal of Advanced Research in Engineering and Technology (IJARET),12(2), 2021, pp. 664-675. http://iaeme.com/Home/issue/IJARET?Volume=12&Issue=2

[14] D. Praveen Kumar, Tarachand Amgoth, Chandra Sekhara Rao Annavarapu, Machine learning algorithms for wireless sensor networks: A survey, *Information Fusion* (2018), doi: https://doi.org/10.1016/j.inffus.2018.09.013

[15] Ali Abdul-hussian Hassan, Wahidah Md Shah, Mohd Fairuz Iskandar Othman,Hayder Abdul Hussien Hassan "Evaluate the performance of K-Means and the fuzzy C-Means algorithms to formation balanced clusters in wireless sensor networks", International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 2, April 2020, pp. 1515~1523, DOI: 10.11591/ijece.v10i2.pp1515-1523.

AUTHOR'S PROFILE

Dr. S. Sukumaran, working as Associate Professor, in Department of Computer Science, Erode Arts and Science College, Erode, Tamilnadu, India. He is a member of Board of Studies and Doctoral Committee Member in various Autonomous Colleges and Universities. In his 35 years of teaching experience, he has supervised more than 55 M.Phil research works, guided 23 Ph.D research works and still continuing. He has presented, published around 80 Research Papers in National, International Conferences and Peer Reviewed Journals. His area of research interest includes Digital Image Processing, Networking, and Data mining.

M. Manoranjani has completed M.C.A degree from Anna University, Coimbatore in 2010. She also awarded M.Phil degree in Computer Science from Bharathiar University in 2013. She has got 8 years of teaching experience. Currently she is pursuing Full-Time Ph.D in Computer Science at Erode Arts and Science College, Erode. Her research interest includes MANET, Wireless Sensor Networks.