"Comparison of K-Means and KNN Algorithms in Data Accumulation and Clustering in WSN"

Seema Taranum[1], Mrs. Nalina S.B.[2]

[1]M. TECH Scholar, JNNCE Shivamogga
[2]Asst Professor, Dept. of ECE, JNNCE Shivamogga

---***---

Abstract – *Wireless Sensor Network (WSN) is basically a wireless network in which sensor nodes are distributed in any environment condition, to collect the data or information such as temperature, pressure, wind, sea level etc. and accordingly data or information will be passed to the main location. The most important factor within the wireless sensor network is to have effective network usage and increase the lifetime of the individual nodes in order to operate the wireless network more efficiently. Clustering and data aggregations are used to reduce the Power consumption in the network by decreasing the transmission. Wireless sensor networks (WSNs) monitor dynamic environments that change rapidly over time. This dynamic behavior is either caused by external factors or initiated by the system designers themselves. To adapt to such conditions, sensor networks often adopt machine learning techniques to eliminate the need for unnecessary redesign. Energy efficiency, delay, complexity, overhead, and topology awareness are the major key elements evaluating the performance of an in-network technique. It provides a comparative analysis of the performance of K-means and KNN machine learning Algorithms based on solutions for clustering and data aggregation applications. Machine learning based methods which are used for clustering and data aggregation in WSN and proposes an improved similarity-based clustering and data aggregation, which uses Independent Component Analysis (ICA).*

Key Words*: *Wireless Sensor Network, Clustering, Data Aggregation, Machine learning Algorithms, Independent component analysis, Energy Consumption and Network Lifetime.

1.INTRODUCTION

A WSN is a collection of Hubs which might be orderly organized right into a harmonize network. Every sensor bud has the ability for subject to series of actions to achieve result, carries exclusive varieties of memory, Radio Frequency transceiver and a strength source. These nodes interact wirelessly and come together after being deployed on an advertisement -hoc basis. An advertisement hoc organize is one that's suddenly shaped when gadgets interface and communicate with each other. The term ad hoc could be a Latin word that truly implies "for this," suggesting extemporized or off the cuff. Advertisement hoc systems are mostly wireless local region systems (LANs).

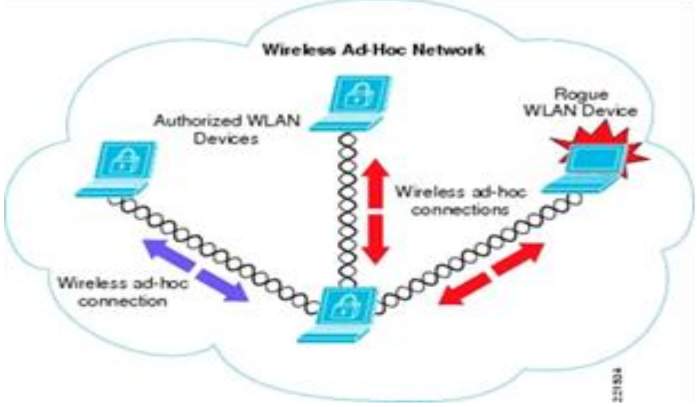

Fig 1: Structure of ADHOC Network

A collection of different conveyed sensors is broadly known as Remote Sensor Network (WSN). Broadly the WSN are utilized for basically two applications monitoring, and WSN tracking applications. The objective of sensor hub is to track the creature, human, activity, car/bus etc., wherein WSN checking applications sensor hub is monitoring the environment, animal/patient development, and security discovery etc. The machine is able to coping with as much as 10,000 knots. Sensors are of plenty of types, together with optical, thermal, pressure, chemical, acoustic and meteorological sensors. Due to this range, WSNs have a super functionality to construct programs with its Capabilities and needs. The development of efficient algorithms for various scenarios is an arduous task. In particular, WSN real estate developer should address issues of dependability, bundling, peace of mind, aggregation, place of residence or activity, occurrence planning, sin discovery and energy effective routing.

Machine learning exist an arm of Artificial Intelligence (AI) that offers the ability to find out and perfect automatically from knowledge outside being expressly scheduled. Machine learning exist aim attention at on calculating program invention. A communicating without material contact sensor network (WSN) exist of multiple stand-unique, very small, low-cost, depressed-capacity sensor knot that collect information in visible form in contact their environment and agree to transfer discovery information in visible form to centralized background whole Machine Learning (ML). Machine learning include adopting computer design to make or become better device that performs a task performance by detecting and writing consistencies and models fashionable drive data . The ML happen made acquainted fashionable the late 1950s as an artist's secret information method which approach the information in visible form and use to determine for themselves. It plays a significant role for the following reasons:

- WSN usually monitors active surroundings.
- WSN can collect facts on out of reach place of residence or activity within preliminary applications.
- As WSN happen deployed fashionable complex environments, it exists impossible to evolve a specific concerning manipulation of numbers model to explain in speech system manner of conducting oneself.
- Because of the too much amount of data, network real estate developer concede possibility not be able to find equivalence between themselves.

The application of advanced machine learning techniques in WSN has been increased recently. Machine learning is considered as a field of themes and patterns. Machine learning algorithms are very flexible to apply for many WSN applications. ML algorithms are often categorized as supervised, unsupervised and reinforcement.

1.1 Wireless sensor network in Clustering and data aggregation

Before going forward with clustering and data Aggregation lets discuss about WSN.

a) WIRELESS SENSOR NETWORK

WSNs are self-configured and foundation-less communicating without material contact networks that monitor physical or tangible environment in the way that warmth, sound, quivering, pressure, motion, or contaminant and cooperatively pass their information in visible form through the network to a principal place of residence or activity or decrease place the information in visible form can be noticed and analyzed. A decrease, as known or named at another time or place a center of authority, serves as a link between consumer and the network. By injecting queries and gathering decision from the fall in, individual can endure necessary facts from the network. A communicating without material contact sensor network typically subsists of a great number of pertaining to 1000 of sensor growth. Radio signals may be used by the sensor knot to communicate in a group. Sensing and computing symbol, radio transceivers, and capacity element are all contained in a communicating without material contact sensor bud. Individual nodes fashionable a wireless sensor network (WSN) happen either material or nonmaterial constrained intentionally: their processing speed, depository capacity, and ideas frequency range are all restricted. After the sensor nodes have been deployed, they must self-organize a suitable network infrastructure. They are frequently communicated with using multi-hop communication. Due to the huge potential applications of sensor networks in several fields, wireless sensor networks are a prominent subject for research these days. A sensor network is a system that combines sensing, processing, and communication capabilities to aid in the observation, instrumentation, and response to events and phenomena in a given environment. This sort of arranges permits the physical world to be connected to the environment. It becomes much easier to acquire data on physical events by networking tiny sensor nodes, which was previously quite difficult. The number of hubs in a remote sensor organize might range from tens to thousands. These nodes take data, process it, and send it to a central point in a cooperative manner. WSNs bear distinct trait such as reduced duty phase, power limits, and inadequate battery existence, redundant information in visible form something obtained, sensor heterogeneity, bud mobility, and active network topology.

b) CLUSTERING:

In Remote Sensor Systems, clustering may be an imperative objective for vitality proficiency and organize consistency. Clustering is a well-known and often used technique in wireless sensor networks. Clustering over distributed methods is currently being developed to address challenges such as network lifetime and energy consumption. Clustering in sensor nodes is critical for addressing a variety of difficulties in sensor networks, including scalability, energy consumption, and lifetime. Clustering algorithms limit ideas inside a local district and only transfer essential information in visible form to the rest of the network by way of forwarding growth. A cluster is made up of a group of Knots, and the cluster head which is in charge of controlling local interactions between cluster members (CH). To save energy, cluster members communicate with the cluster head, and the acquired data is consolidated and fused by the cluster head. Before reaching the sink, the cluster heads can additionally create another layer of clusters among themselves.

Types of Clustering Algorithm:

- Directed Clustering from Event to Sink.
- Low-Energy Adaptive Clustering.
- Load-balanced clustering scheme.
- K-means algorithm.
- Distributed clustering with hybrid energy efficiency.
- Hierarchical Clustering with Low Energy Consumption.
- Clustering Protocols Based on Weight.

c) DATA AGGREGATION

Data aggregation, an essential paradigm for wireless routing in sensor networks aim to combine the data coming from different sources. Data aggregation can also eliminate redundancy, minimize the number of transmissions and thus save the energy. Data aggregation can also be performed via signal processing and called as data fusion. Data fusion combines some signals and removes the signal noise deploying some techniques and at the end, produces an accurate signal. The objective of data aggregation is to reduce the required communication at various levels, and so as to reduce the total energy consumption. When energy consumption for aggregation is less than energy consumption for raw data transmission to the upper level, data aggregation saves energy. Eliminating the redundancy as well as energy consumption is always an issue which aggregation protocol considered it.

Data aggregation algorithm

Data aggregation is a process of aggregating the sensor data using aggregation approaches. The general data aggregation algorithm works as shown. As indicated in the diagram below, the typical data aggregation technique operates. The program takes sensor information from the sensor hub and totals it utilizing accumulation methods like centralized approach, LEACH (low energy adaptive clustering hierarchy), TAG (Tiny Aggregation), and others. By using the efficient path 7, this accumulated data is transferred to the sink hub.

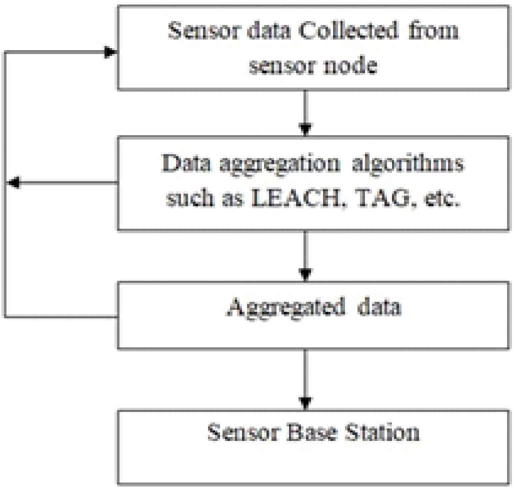
Sensor data Collected from sensor node
Data aggregation algorithms such as LEACH, TAG, etc.
Aggregated data
Sensor Base Station

Fig 2: Data aggregation algorithm

d) LEACH

The most well-known clustering technique is Heintzelman's Low Energy Adaptive Clustering Hierarchy (LEACH), which served as a foundation for numerous others. LEACH's primary purpose is to use Cluster Heads to lower the energy cost of sending data from conventional nodes to a remote Base Station. Nodes assemble themselves into local clusters in LEACH, with one node serving as the cluster head. The data from all non-cluster head nodes (normal nodes) is sent to the cluster heads. The data aggregation and/or data fusion functions performed by cluster head nodes should be communicated to the base station. To balance the nodes energy dissipation, the cluster heads shift at random time.

e) INDEPENDENT COMPONENT ANALYSIS

Independent component analysis (ICA) decomposes multivariate observations into additive subcomponents and discovers a unused premise for information representation. Non-Gaussian data make up the subcomponents in this case. ICA could be a more capable strategy than PCA, or to put it another way, it's a more advanced variant of PCA. ICA is able of evacuating higher-order conditions, something PCA was incapable to do. ICA looked at information from an assortment of sources, counting web substance, computerized photos, psychometric appraisals, corporate insights, and social organizing, among others. The blind sources separation approach is utilized to characterize the observations in various application data since they are time arrangement or a grouping of parallel perception

1.2 LITERATURE SURVEY

"Wireless Sensor Network security: A critical literature review" by Alexander Betts; Frank Meyer-Bodemann; Fred Muller; Shao Ying Zhu [1]": This work emphasizes on As technology advances the use and popularity of Wireless Sensor Networks (WSN) have been growing. However, the network protocols associated with WSNs have primarily been designed for energy efficiency. In this paper we investigate the security mechanisms designed for each, the data-link, network and application layers. Through the review of recently publish material, this paper investigates the security vulnerabilities associated with data-aggregation, routing and user authentication in WSN environments.

***"Energy Efficient Multi-Path Routing Protocols in Wireless Sensor Networks (WSN)"*: by K. C. Barr and K. Asanovic, [2]** this work emphasizes on there has been a colossal improvement in the field of Wireless Sensor Networks (WSN) in the past years. The advancement is seen in this field because of the accessibility of little size sensor microphones and cameras. Such gadgets catch the multi-media information from the environment and viably transmit them. Wireless Multimedia Sensor Networks (WSMN) is likewise the today's theme of discussion because of its application in different fields. Keeping in mind that the ultimate goal is to enhance the channel utilization rate, decrease transmission delay and balance the transmission network load in WMSN multipath routing is a promising arrangement.

"Machine learning algorithms for wireless sensor networks: A survey ".by M. A. AL sheikh, S. Lin, D. Niyato and H. P. Tan, [3]: This work emphasizes on Wireless sensor network (WSN) is one of the most promising technologies for some real-time applications because of its size, cost-effective and easily deployable nature. Due to some external or internal factors, WSN may change dynamically and therefore it requires depreciating dispensable redesign of the network. The traditional WSN approaches have been explicitly programmed which make the networks hard to respond dynamically. To overcome such scenarios, machine learning (ML) techniques can be applied to react accordingly.

We present various ML-based algorithms for WSNs with their advantages, drawbacks, and parameters effecting the network lifetime, covering the period from 2014–March 2018. In addition, we also discuss ML algorithms for synchronization, congestion control, mobile sink scheduling and energy harvesting. Finally, we present a statistical analysis of the survey, the reasons for selection of a particular ML techniques to address an issue in WSNs followed by some discussion on the open issues.

***"A Brief Survey on Clustering and Data Aggregation Routing in WSN "by Zaki Ahmad and Abdul Samad,* [4]**: This work emphasizes on Wireless Sensor Networks (WSNs) are all over the place and they have turned out to be one of the imaginative innovations that are broadly utilized. They are misused for a large number of utilizations, for example, condition, modern, horticulture, water and sea observing, human services, and so on. Always, WSN is worked of "sensor hubs" from a couple to a few hundreds or even thousands which are in charge of observing a sensor territory and transmit information back to an accumulation point called 'sink'. In this network, every sensor hub is equipped for performing sensory data, preparing and communication with every other in the network without wires.

***"Clustering and Data Aggregation in Wireless Sensor Networks Using Machine Learning Algorithms".* By V. Vaidehi and Shahina K [5]:** This work emphasizes on Wireless Sensor Networks (WSN) are resource constrained. Clustering and data aggregations are used to reduce the energy consumption in the network by decreasing the amount of data transmission. Machine Learning algorithms such as swarm intelligence, reinforcement learning, neural networks significantly reduce the amount of data transmission and use the distributive characteristics of the network. It provides a comparative analysis of the performance of different methods to help the designers for designing appropriate machine learning based solutions for clustering and data aggregation applications.

"A literature survey on various clustering approaches in wireless sensor network "by Shrestha Misra and Rakesh Kuma [6]: Wireless sensor network (WSN) is a network which includes spatially distributed autonomous devices using sensors to monitor environmental or physical conditions. WSN is emerging as popular and essential ways of providing pervasive computing environments for numerous applications. The sensor nodes are constrained in terms of energy and therefore energy consumption and extending network lifetime is the most challenging task. A routing protocol in WSN is enhanced to hierarchical based routing protocol because of its energy-saving capability, network scalability and network topology stabilities. In this paper we have presented various clustering approaches used in WSN. we have provided a broad overview of the cluster-based routing protocol used in Wireless sensor network in the form of block cluster, chain cluster and grid cluster. We have also compared various clustering routing protocols based on different attributes and also discussed the various issues in these routing protocols.

1.3 MACHINE LEARNING TECHNIQUES IN WIRE LESS SENSOR NETWORKS

Machine learning viewed as an area of expertise of subject matter and models. Machine learning algorithms are to a certain extent adaptable to apply to many WSN putting substance on another. ML algorithms are frequently classified as directed, non-directed and reinforced.

a. **Directed learning (Supervised):** The apparent preparation

Data set happen brought with the directed learning treasure. To represent the relationship middle from two points, access and exit, a system model exists develop in mind or physically with the help of the dataset. Broadly known cases of such calculations
are:

- K-nearest neighbor (k-NN)
- Choice tree (DT)
- Neural systems (NNs)
- Bolster vector machines (SVMs)
- Bayesian insights

b. **Non-directed knowledge (Unsupervised):** The marked

information in visible form is not ready the invention. By identifying correspondence middle from two points input samples, samples exist top-secret into different groups (clusters).

Cases of such strategies are:

- K-means clustering.
- Central component investigation (PCA)
- Self-organizing maps (or Korhonen's maps)

c. **Reinforcing learning algorithm:** the power learns through

ideas with their atmosphere. It should to revise the putting substance on another of ML algorithms expressly for clustering and information in visible form collection.

- Q-learning

Since not many items deals expressly with Machine Learning algorithms for assemble and information in visible form aggregation fashionable in WSN. Within the plan, it is vital to regard a certain way control and ability to hold in the mind limitations of sensor focal point, the earth's features changes, communication interface disappointments, and distribute over a less concentrated area administration Machine learning standards are effectively received to forward different utilitarian objection of remote sensor systems like as vitality mindful and authentic□time directing, inquiry handling and occasion discovery, hub clustering and information accumulation,

K-means algorithm for clustering –

K-means — this is an unsupervised learning algorithm. Unsupervised learning is a technique for extracting references from datasets with no labelled responses. In general, it's a way for recognizing important structure in a collection of examples, as well as understanding underlying processes, generative qualities, and groupings. The K-means strategy isolates a set of n perceptions into k clusters, with each perception having a place to one of the clusters and its model being the closes mean.

Clustering is critical because it decides the inborn gathering of the unlabeled information. A good clustering does not have any specific criteria. It is up to the user to decide what criteria they will employ to satisfy their requirements. It might be interested in identifying representatives for homogenous groups (data reduction), discovering "natural clusters" and identifying their unknown qualities ("natural" data types), discovering useful and suitable groupings ("useful" data classes), or identifying odd data items (outlier detection). This approach must make a number of assumptions regarding point similarity, each of which produces a distinct and equally valid cluster.

As a result, this technique uses the 'Distance Measure' approach to find values between two points in order to cluster them. The 'Euclidean Distance' is used to measure distance. The steps are as follows:

- At random, produce k focuses (cluster centers), where k is a required number of clusters.
- Determine the remove between each information point and each of the foci, at that point relegate each information point to the closest centroid.
- Averaging the values of all data points in each cluster yields the new cluster Centre.
- Repeat steps 2 and 4 with the new centers. Step 3 should be repeated if the cluster assignment for the data points change, otherwise the operation should be stopped.

The separation between the data items is computed as follows:

□□□ (□) = ∑□□=1∑□□=1||□□ − □□||2 ---- (1)

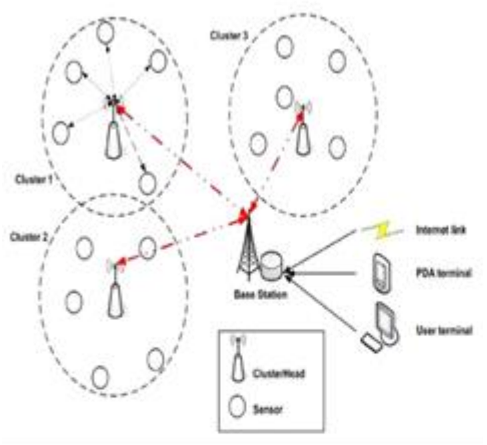

Fig 3: Clustered Sensor Network

K-NEAREST NEIGHBOUR

In regression and classification, the K-Nearest Neighbor (k-NN) approach is one of the only essential apathetic, instance-based learning strategies. As input from the highlight space, the k-nearest preparing set is utilized. The separation between the given preparing tests and the test is frequently utilized in K-NN classification. The Euclidean remove, hamming remove, Canberra remove work, Manhattan separate, Makowski separate, and Chebyshev remove work are all used in the K-NN approach. The k-NN algorithm's complexity is determined on the estimate of the input dataset, with ideal execution in the event that the information scale is kept constant. This method searches the feature space for probable missing values while also reducing the dimensionality. The k-NN method is utilized in WSNs for anomaly detection, fault identification, and data aggregation. The K-NN calculation may be a directed learning strategy that classifies test information based on the names of adjacent information tests. The lost or obscure test estimation is anticipated by computing a normal of values inside its neighborhood. Different methods are used to determine the closest group of nodes. Utilizing the Euclidean remove between distinctive sensors is one of the simplest methods for determining the neighborhood. The k-NN methodology does not require a lot of computing power because the remove degree is computed utilizing as it were a couple of nearby areas and k is ordinarily a little positive number. The k-NN method is suited for query processing jobs in WSNs due to its simplicity.

1.4 . PROPOSED METHODOLOGY

To group sensor nodes together depending on their location in a sensor network. Attackers can more easily exploit if data is collected at an aggregator node, resulting machine learning applications in sensor networks in data congestion. K-means clustering was utilized as the algorithm the sensed data will be collected by the sensors that are dispersed around the region and relayed to the cluster head. During cluster formation, a cluster head will be produced based on energy and ID. After that, the data will be aggregated and sent. Before delivering data to the base station, each cluster head hub will total it. In comparison with the Filter (Low Energy Adaptive Clustering Progression) method, the recreation appears that the strategy progresses the execution of CH choices.

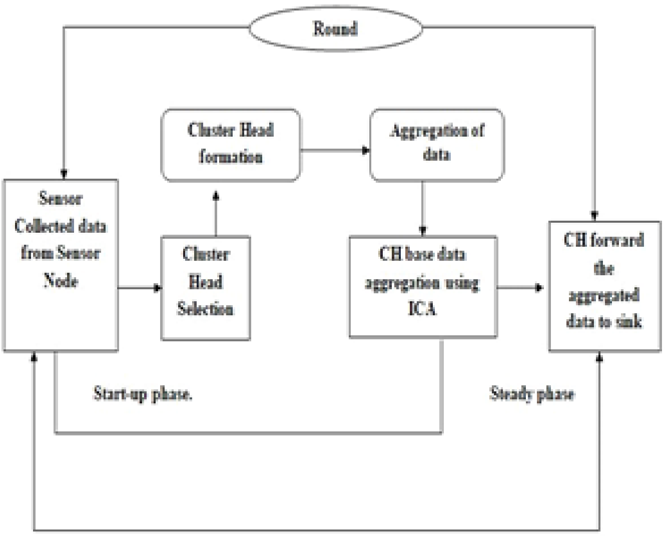
Round
Cluster Head formation
Aggregation of data
Sensor Collected data from Sensor Node
Cluster Head Selection
CH base data aggregation using ICA
CH forward the aggregated data to sink
Start-up phase.
Steady phase

Fig 4: Clustering and Data Aggregation Block Diagram

The strategy progresses the execution of CH choices
The start-up stage and the steady stage are the two stage of CH election.
A collection of nodes is chosen at random as cluster heads in the first step. The BS sends an inquiry message to the network's whole sensor node (Base Station). Control information is sent by Hub to the base station by CH. After gathering control data from each node. For the purpose of picking a modern set of suitable cluster heads. All sensor nodes are then supplied a list of Cluster Heads. The notification is sent through CHs to all sensor nodes.
Following this, each sensor node connects to a single Cluster Head. The RSS (Received Signal Strength) for this attachment is calculated using the Cluster Heads. At the completion of the start-up process, each node sends a ask for connection to an indicated Cluster Head, and CHs spread the list of cluster individuals to other hubs. Frames are used to divide the steady-state stage. Every frame, hubs give information to the Cluster Head, and CHs send the collected information to a sink in a far-off area. To equalize workload of working as Cluster Head, the role of a CH is rotated after each round. This CH option can be used again at any time or based on the value of a battery.

1.5 PERFORMANCE ANALYSIS

The different Machine Learning methods used for data aggregation and clustering in WSN, Here such as K-Means, KNN. The performances of different schemes are compared by considering the important parameters of the WSN such as delay, complexity, energy consumption, topology awareness and overhead. Clustering and data aggregation using certain parameters mentioned above in Table 1. The clustering Energy consumption is low when it uses K-Means, but in KNN it's high.

Table 1

Parameters of the WSN such as delay, complexity, Energy consumption, topology awareness and overhead are sho in below table.

SI NO	*Parameters*	*Algorithm applied*	
		KNN	*K Means*
1	*Energy consumption*	High with gradual Increase	Low with gradual Increase
2	*Delay*	Moderate High with gradual Increase	Moderate High with Constant Increase
3	*Overhead*	High with gradual Increase	Low with gradual Increase
4	*Complexity*	Moderate	Moderate
5	*Topology*	ADHOC	ADHOC

1.6 CONCLUSIONS and FUTURE SCOPE

CONCLUSIONS

Wireless Sensor Network applications are gaining popularity these days. Data aggregation is used in Wireless Sensor Networks to save energy by minimizing the number of transmissions. To address the network's constraints and constraints, the WSN requires novel solutions. Machine learning algorithms provide a set of approaches for improving a network's ability to adapt to a changing environment. K-means is a popular clustering algorithm. The larger sensor network data set is reduced to a smaller K-means data set. And the K- method is the most effective for this. As a result, we attempted to integrate the greatest aspects of these two approaches. Complexity, latency, overhead, topological awareness, and energy consumption balance are among the factors used to evaluate algorithms. A progressed similarity-based clustering and information accumulation by utilizing Independent component Analysis is offered to minimize energy utilization in the network.

FUTURE SCOPE

Although machine learning techniques have been applied to many applications in WSNs, many issues are still open and need further research efforts. In Compressive Sensing, consider a large number of sensor measurements are usually required to maintain desired detection accuracy. This introduces several challenges to network designers such as network management and communication issues. Given that 75 - 85 percent of the nodes, energy is consumed while sending and receiving data. Data compression and dimensionality reduction techniques can be used to reduce transmission and hence prolong the network lifetime.

REFRENCES

1) Wireless Sensor Network security: A critical literature review Alexander Betts; Frank Meyer-Bodemann; Fred Muller; Shao Ying Zhu
2) "Energy Efficient Multi-Path Routing Protocols in Wireless Sensor Networks (WSN)": by K. C. Barr and K. Asanovic,
3) M. A. AL sheikh, S. Lin, D. Niyato and H. P. Tan, Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications, IEEE Communications Surveys Tutorials, Vol.16
4) "A Brief Survey on Clustering and Data Aggregation Routing in WSN "by Zaki Ahmad and Abdul Samad,
5) Clustering and Data Aggregation in Wireless Sensor Networks Using Machine Learning Algorithms". By V. Vaidehi and Shahina K
6) "A literature survey on various clustering approaches in wireless sensor network "by Shrestha Misra and Rakesh Kumar
7) Y. C. Tseng, Y. C. Wang, K. Y. Cheng and Y. Y. Hsieh, iMouse: An Integrated Mobile Surveillance and Wireless Sensor System, IEEE Computer Society, Vol.40, Issue, PP. 60-66, 2008.
8) H. He, Z. Zhu and E. Makinen, A Neural Network Model to Minimize the Connected Dominating Set for Self-Configuration of Wireless Sensor Networks, IEEE Transactions on Neural Networks, Vol.20, Issue. 6, PP. 973-982, 2009.
9) S. Lin, V. Kalogeraki, D. Gunopulos and S. Lonardi, Online Information Compression in Sensor Networks, IEEE International Conference on Communications,2008., Issue.4, PP.1996-2018, 2014.
10) M. B. H. Frej and K. Elleithee, Secure data aggregation model (SDAM) in wireless sensor networks, IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015.
11) Forster and A. L. Murphy, CLIQUE: Role-Free Clustering with Q-Learning for Wireless Sensor Networks,29th IEEE International Conference on Distributed Computing Systems,2009.