Machine Learning Based Clustering Techniques for Wireless Sensor Networks: An Overview

Reem Abu Taleb
Department of Management Information Systems
Business Faculty, Al-Balqa' Applied University

Abstract - Wireless sensor networks are expected to operate in harsh and dynamic environments without intervention. As a result, these networks need to have the ability to adapt to changes in the environment and in their topologies in order to maintain high performance levels. Energy consumption is one the major resources affecting the operation of wireless sensor networks. Thus, various clustering techniques were proposed in order to decrease the number of intermediate node data packets need to go through in order to reach the base station. In this paper, an overview of machine learning based clustering techniques is presented because, providing intelligent rather than traditional clustering technique will help in prolonging sensor nodes and sensor networks lifetime and increases energy efficiency.

Keywords: Wireless Sensor Networks, Machine Learning, Clustering, performance

I. INTRODUCTION

Wireless sensor networks (WSNs) consists of small sensor nodes that might be randomly scattered in the area of interest and communicate via wireless communication. Furthermore, these networks are expected to operate for long periods of time without any human intervention since they might be deployed in hostile environments. As a result, sensor nodes are expected to provide large amounts of information regarding the area of deployment and the phenomena being studied. When a sensor node obtains the sensed data, it will report it to the base station in order to be further analyzed [1][2].

Therefore, WSNs have been used in various applications in many different areas. To name a few, these networks can be used in military applications for surveillance and tracking. Additionally, they can be used in agricultural applications for monitoring temperature and pressure. Furthermore, WSNs can be used in healthcare systems to provide real time monitoring of patients. Also, WSNs can be used to track vehicles in order to prevent traffic congestion [2].

Since sensor nodes communicate via wireless communications and are deployed in large numbers, they are expected to be cheap with limited resources such as limited battery power, limited memory capacity and limited processing capability. Also, sensor nodes are expected to be able to manage and reconfigure themselves since they are expected to operate in an unattended manner for long time periods [3].

As a result, many research papers have proposed different techniques to enhance the performance of WSNs and solve the main challenges for these networks. The research proposed in [3] listed the main challenges to be addressed in the field of WSNs, some of them are listed below:

- *Self-Management*

  Sensor nodes are expected to manage network configuration and adapt to changes that might appear in the network topology.

- *Limited memory and storage space*

  Since sensor nodes are cheap and have limited hardware resources, node have to use these resources wisely. As a result, the techniques and the software used by sensor nodes must be light weight to avoid causing buffer

overflow. Also, sensed data must not be saved in memory for long period of time to manage them memory wisely and make sure that the sensed data does not get old and obsolete.

- *Fault Tolerance*

  WSNs must be able to adapt to node failures which results in changes in connectivity and affects the performance. Thus, these networks must be able to remain functional with the presence of failures. In other words, the network is expected to maintain its functionality on the expense of some performance degradation.

- *Energy*

  As mentioned before, sensor nodes are battery operated. As a result, their lifetime is restricted to the lifetime of their batteries. Consequently, sensor nodes are expected to wisely use their energy sources and avoid energy depletion in order to extend the network lifetime and maintain high performance levels.

  Many research papers have addressed the limited energy challenge and have proposed various techniques in improve the energy efficiency of WSNs. Dividing sensor nodes into clusters is one of the major techniques that was proposed in order to limit the number of node a sensor node communicates with. Thus, energy efficiency can be achieved. In addition, different methods were proposed in order to achieve clustering in an effective manner.

  The rest of this paper proceeds as follows; in section 2 the importance of clustering is discussed. In section 3 the main categories of machine learning clustering techniques are presented. After that, the techniques falling under each category are presented. Finally, the paper is concluded in section 5.

II. IMPORTANCE OF CLUSTERING

As mentioned before clustering can be adopted to reduce energy consumption of sensor nodes and achieve energy efficiency. According to [4], traditional clustering methods address the number of clusters to be constructed in order to enhance the performance of the network. Additionally, the number of nodes to be included within a cluster is another issue to be addressed. Finally, these techniques aim to address cluster head and replacement issues. Because traditional clustering methods rely on two parameters to select the cluster head. The first parameter is the energy level of the node while the second parameter is the distance of the node from the base station. As a result of relying on these two parameters, cluster heads will be overcrowded in dense networks. On the other hand, very small number of cluster heads may exist in sparse networks [5].

As a result, Machine learning (ML) based clustering techniques were proposed in order to overcome the problems of the traditional clustering methods. ML based clustering techniques aim to find the optimal number of cluster heads to be deployed [5]. Also, decide on cluster head substitution method. As a result, they make sure that the network is not crowded with cluster heads. Moreover, clustering advantages can be further developed using ML techniques because these methods help to form clusters and to select cluster heads in an energy efficient fashion and avoid consuming most of sensor nodes energy in clustering and cluster head selection stages [6].

The main focus of this paper is to review and study ML clustering techniques. As a result, in the remaining sections of this paper various ML clustering categories and techniques are discussed and presented.

III. MACHINE LEARNING TECHNIQUES

According to [7][8][9][10] machine learning clustering techniques can be divided into three categories namely, supervised, unsupervised and reinforcement learning. On the other hand, the research presented in [11] classified ML clustering techniques into four categories based on intended structure of the model. As a result, they can be classified into supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Additionally, the research presented in [12] has introduced a new category named evolutionary computation. Thus, ML clustering techniques can be classified into the following categories; supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and evolutionary computation.

As a result, in this paper ML clustering techniques are classified into five categories, namely supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and evolutionary computation, in

order to provide a more general view of these techniques. Also, several algorithms fall under each category which will be further discussed in this paper.

Supervised learning is based on learning by example based on the relationship between input and output parameters [7]. Decision trees, K- Nearest Neighbor, Neural networks, Support vector machine and Bayesian Network are the algorithm types that fall under this category [6].

In unsupervised learning the algorithm does not provide output vectors or labels. However, the data sets to be used can be classified in order to find similarities between them [7]. Principal component analysis and K-means clustering and the most two distinguished types of algorithms that can be classified under this category [10]. On the other hand, semi-supervised learning is a hybrid category that aims to inherit the advantages of supervised and unsupervised learning while reducing their disadvantages [11]. According to [13], semi-supervised learning can be further classified into two sub categories; transductive learning and inductive learning.

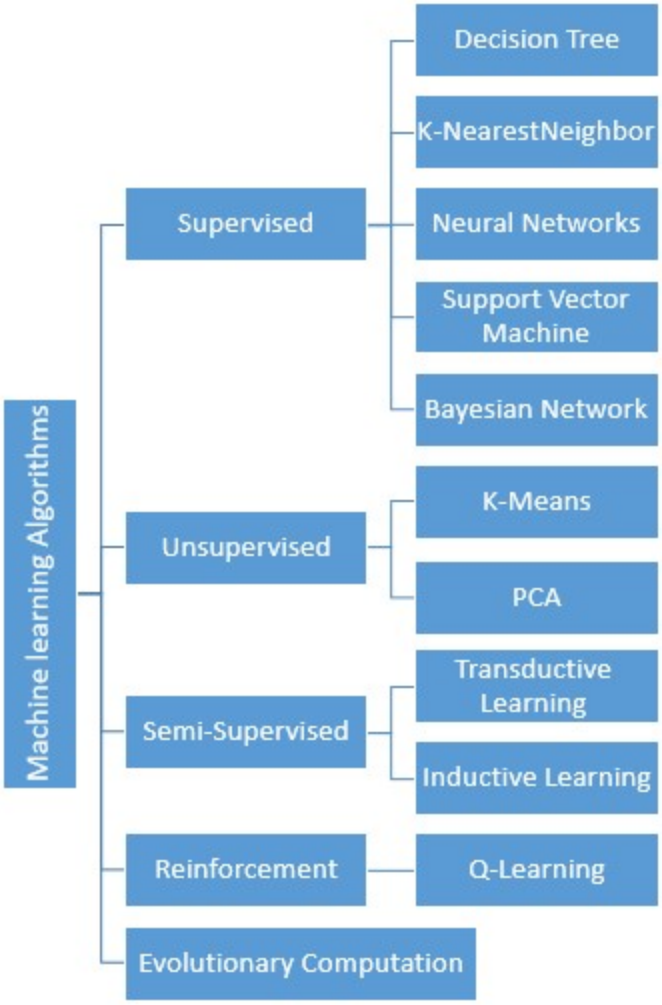

Figure 1: Machine Learning Techniques

Additionally, reinforcement learning is based on giving the agent the ability to interact with its environment in order to learn from it. In WSN sensor nodes can learn to capture the best measurements in order to maximise their advantage [7]. According to [10], Q-learning is the most distinguished reinforcement learning algorithm that can be used in the field of WSN in order to solve routing problems.

Finally, Evolutionary computation is based on various optimization techniques that are derived from biological evolution and nature. Thus, several iterations must be applied in order to obtain the solution. Algorithm in this

category are used to solve several issues of WSNs such as coverage problem, localization, target tracking, routing, and mobile sink issues. Ant colony optimization, genetic programming, genetic algorithms, evolutionary programming, evolutionary algorithms, artificial immune systems, artificial bee colonies, particle swarm optimization, and firefly algorithms are the algorithms and techniques that fall under this category [12]. Figure 1 shows the classification of ML algorithms that can be used to achieve clustering in WSNs

IV. MACHINE LEARNING BASEED CLUSTERING TECHNIQUES

In this section several techniques falling under the categories presented in figure 1 will be presented and discussed.

4.1. Supervised Learning

As mentioned in section 3, supervised learning algorithms can be divided into several categories namely, Decision trees, K- Nearest Neighbor, Neural networks, Support vector machine and Bayesian Network.

4.1.1. Decision Trees

The authors of [14] proposed using decision trees in order to solve cluster head selection problem. The proposed technique is based on dividing the network into non-overlapping clusters and selecting the node with the highest energy level to be the cluster head for each cluster using decision tree algorithm. Battery level, distance from the center of the cluster, vulnerabilities indications and the degree of mobility, for the input vector iteration of the decision tree.

A cluster head election scheme that aims to reduce the intra-cluster communication distance in order to reduce the total energy consumption in the whole network was proposes in [15]. Here, cluster head selection is based to the amount of residual energy and the cost required for intra cluster communication which is also dependent several factors such as the size of the cluster and the distance between nodes within the cluster.

4.1.2. K-Nearest Neighbor

Collaborative signal processing was used in the research proposed in [16] in order to achieve distributed detection and tracking of a single target in wireless sensor networks. Furthermore, using k-nearest neighbor and support vector machines were adopted in order to provide the ability to track multiple objects. As a result, massive amounts of information has to be collect from the sensor nodes. The main components of the system are event detection, estimation, prediction and classification.

The authors of [17] proposed a clustering algorithm that combines hierarchical and distance based clustering approaches. Additionally, the proposed algorithm is based on using K-nearest neighbor technique in order to divide the network into clusters. First, the algorithm the location of every sensor mode is collected by the base station. Then, for each sensor a collection of k-nearest neighbor is calculated. After that, the pairwise distances between a sensor node and its top k-nearest neighbors is calculated. Finally, all pairs with distances below a specified threshold are merged.

4.1.3. Neural Networks

A self-organizing clustering method based on neural networks was proposed in [18]. The method is based on using neural networks in order to specify the minimum connected dominating set. After that, cluster heads are selected for the nodes that are part of the connected dominating set. Then, the method is based on deploying a mobile sink that will traverse the connected dominating set nodes i.e. the cluster heads in order to collect data from them.

Moreover, the research proposed in [5], presented an algorithm that is based on using machine learning techniques in order to divide the network into clusters and apply data aggregation to improve the energy efficiency of the network. Neural networks is used to form the clusters within the network. In their research, cluster head selection process is based on the architecture of the neural network where the residual energy and the distance from the base station and the main properties of every sensor node to be used as an input for the neural network. As a result, every node is evaluated according to the evaluation function that used the previously mentioned properties of the nodes. After that, competition takes place in the hidden layers of the neural network and the process is repeated

until cluster heads are selected. Note that, cluster heads selection phase is executed periodically in order to avoid depleting the energy of the nodes that were selected to act as cluster heads.

4.1.4. Support Vector Machine

Support vector machines can be defined as a machine learning technique that has the ability to learn how to classify data points based on labeled training samples. The functionality of support vector machine is based on dividing the space into parts that are separated by separation gaps. After that, the new inputs or readings will be classified according to these gaps [11]. The authors of [19] proposed using least square support vector machines in order to estimate the clusters of nodes in wireless sensor networks. The technique is based on using mixtures of kernels and images representing the distribution of energy in the field. Also, in [20] the use of support vector machine was proposed in order to achieve clustering and reduce the amount of energy consumed.

4.1.5. Bayesian Network

The authors of [21] proposed dividing the network into two level clustering hierarchy that is based on a two level Bayesian model to collect and predict data of sensor nodes. The network architecture is based on dividing the network into clusters where initial cluster heads are selected based on the number of neighbors. In other words, nodes with the highest number of neighboring nodes are initially selected to be the cluster heads thus, the cluster head will be in the center of the cluster. After that, within each cluster the node with the highest level of energy will be selected as the cluster head. Then, every sensor node within a cluster will send to the cluster head the Bayesian model parameters deduced from old or historical data collected from the environment.

A new Bayesian clustering algorithm was proposed in [22]. The proposed algorithm is based on hidden Markov random fields in order to be able to model the locations and the neighborhood relationships for cluster members. Also, Markov chain Monte Carlo procedure was adopted to implement the proposed work efficiently. Furthermore, this procedure can help to control the number of clusters formed.

4.2. Unsupervised Learning

According to [7], the algorithm does not provide output vectors or labels. Furthermore, based on the similarities found among them datasets can be classified. As a result, unsupervised learning algorithms can be used for clustering and data aggregation in wireless sensor networks. Principle component analysis and K-means clustering are two important types of algorithms that fall under this category.

The research proposed in [23] proposed dividing nodes in to clusters based on sensed data correlation rather than relying on energy level or locations of sensor nodes. As a result, the network is divided into separate clusters where nodes reading or data collected is correlated within the each cluster. First, the principle component analysis based clustering estimates covariance matric. After that, an eigenvector covariance matrix is obtained. Then, the number of optimal clusters that can be formed in estimated. Consequently, k-means clustering is used to implement clustering. Finally, the network is divided into the specified number of clusters.

Another algorithm that combines the use of distributed eigenvector computation and distributed k-means clustering was proposed in [24] in order to reduce the amount of data transmitted and avoid congestion. As a result, the technique is based on using power iteration scheme in order to compute the eigenvector of the graph Laplacian. After that, k-means clustering is applied on the eigenvector.

Moreover, the research proposed in [25] combines the use of k-means algorithm, kohenon self-organizing maps with neural networks conscience function. At the beginning, the K-mean clustering algorithm is applied in order to train the Kohenon self-organizing map using a low dimensional representation of the input space. After that, two parameters, namely energy level and network space, are used by the kohenon self-organizing map in order to group nodes into clusters based on the minimum distance. Then cluster heads are selected based on the remaining energy, level, the distance from the center of the cluster and the shortest distance to the base station. After selecting the cluster head, they will be responsible for communication data collected from sensor nodes within their clusters to the base station.

4.3. Semi-Supervised Learning

The algorithms that fall under this category combine the characteristics of supervised and unsupervised learning algorithm. Thus, they aim to maximize the advantages of these two categories and minimize the disadvantages of

them [11]. According to [13], semi-supervised learning algorithms can be divided into two categories namely; inductive learning and transductive learning. In inductive learning algorithms use a function that is expected to be a good predictor on the data collected in the future. On the other hand, transductive learning algorithms are used to predict the exact labels to be used for an unlabeled data set.

The research proposed in [26] is based on collecting data from the environment then use a semi-supervised learning algorithm to predict future actions. The proposed algorithm is based on dividing the network into clusters and on using a rule based semi-supervised classification model in order to detect abnormal behavior within the clusters. Also, using this classification model helps in categorizing the clusters into different groups namely; high active, medium active mad low active respectively.

4.4. Reinforcement Learning

According to [13], reinforcement learning algorithm are based on continuous learning that can be achieved by directly interacting with the environment and gathering the required information in order to take certain actions. The main goal of reinforcement learning algorithms is to determine the optimal result so that the performance can be maximized.

A role free clustering technique that is based on Q-learning was proposed in [27]. Q-learning is a reinforcement learning technique where agents take actions and receive rewards according the actions they took. In other words, every action is assigned a Q-value that represents the goodness of it. After that, every agent will select and execute an action. Consequently, agents will receive the Q-value that was assigned to the executed action [28].

The research proposed in [27] is based on Q-learning. As a result, nodes have the ability to decide whether they can act as a cluster head for each packet in an independent manner. Thus, there is no real assignment of a cluster head within the cluster. Therefore, every node has an independent learning agent and a set actions representing routing options via different neighboring nodes. As a result, the nodes is selected to be the node with the lowest cost to all sink nodes.

4.5. Evolutionary Computation

Evolutionary computation is part of artificial intelligence that is based on combining several optimization techniques as a problem solving paradigm that uses computational models derived from nature and biological evolution [13].

The research presented in [29] proposed a differential evolution based algorithm that uses a search procedure named diversified vicinity procedure in order to achieve a tradeoff between the energy consumption and the delay incurred when forwarding packets. Nonlinear programming formulation was used to model the problem of data delivery and energy consumption. After that, the generated formulas were solve using differential evolution algorithm.

V. CONCLUSIONS AND FUTURE WORK

In this paper, the importance and characteristics of the wireless sensor networks were discussed. After that, the challenges facing wireless sensor networks were addressed and highlighted. Then, the importance of clustering in wireless sensor networks was discussed. Furthermore, a categorization of machine learning clustering technique was provided. Finally, the techniques falling under each category were briefly discussed for reader convenience.

This paper can be further extended in future to include more techniques and provide applications were each category is useful. Additionally, a comparison, in terms of the methodology and the performance, of the techniques falling under each category may be provided in order to provide a taxonomy that might help researcher to make it easier to decide on the technique to be used based on the problem being addressed.

REFERENCES

[1] Kandris, D.; Nakas, C.; Vomvas, D.; Koulouras, G.,“Applications of Wireless Sensor Networks: An Up-to-Date Survey”,. *Applied. System. Innovation.* **2020**, *3*, 14.

[2] S. R. JinoRamson and D. J. Moni, "Applications of wireless sensor networks — A survey," 2017 International Conference on Innovations in Electrical, Electronics, Instrumentation and Media Technology (ICEEIMT), Coimbatore, 2017, pp. 325-329, doi: 10.1109/ICIEEIMT.2017.8116858.

[3] S. Sharma, R. K. Bansal and S. Bansal, "Issues and Challenges in Wireless Sensor Networks", 2013 International Conference on Machine Intelligence and Research Advancement, Katra, 2013, pp. 58-62, doi: 10.1109/ICMIRA.2013.18.

[4] Nandini. S. Patil, P. R. Patil "Data Aggregation in Wireless Sensor Network", IEEE International Conference on Computational Intelligence and Computing Research, 2010.

[5] SangeetaKumari, RajatUpdhyay , VivekDeshapande," Energy Efficient Clustering and Data Aggregation Protocol using Machine Learning in Wireless Sensor Networks", International Journal of Grid and Distributed Computing,Vol. 13, No. 2, pp. 426–439, 2020.

[6] chander, B., Kumar.B, P., & ., K. (2018).," A Analysis of Machine Learning in Wireless Sensor Network. *International Journal of Engineering & Technology", 7*(4.6), 185-192. doi:http://dx.doi.org/10.14419/ijet.v7i4.6.20460

[7] S. K. and V. Vaidehi, "Clustering and Data Aggregation in Wireless Sensor Networks Using Machine Learning Algorithms," 2018 International Conference on Recent Trends in Advance Computing (ICRTAC), Chennai, India, 2018, pp. 109-115, doi: 10.1109/ICRTAC.2018.8679318.

[8] V. V. Ghate and V. Vijayakumar, "Machine learning for data aggregation in WSN: A survey", *Int. J. Pure Appl. Math.*, vol. 118, no. 24, pp. 1-12, 2018.

[9] Vijayakumar V. (2019)," Application of Machine Learning in Wireless Sensor Network", In: Shen X., Lin X., Zhang K. (eds) Encyclopedia of Wireless Networks. Springer, Cham. https://doi.org/10.1007/978-3-319-32903-1_282-1

[10] Khan, Zaki&Samad, Abdus. (2017). ,"A Study of Machine Learning in Wireless Sensor Network",. International Journal of Computer Networks And Applications. 4. 10.22247/ijcna/2017/49122.

[11] Abu Alsheikh, Mohammad & Lin, Shaowei&Niyato, Dusit& Tan, Hwee Pink. (2014).," Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications", IEEE Communications Surveys & Tutorials. 16. 10.1109/COMST.2014.2320099.

[12] Srilakshmi, N. and A. K. Sangaiah. "Selection of Machine Learning Techniques for Network Lifetime Parameters and Synchronization Issues in Wireless Networks." *„J. Inf. Process. Syst.* 15 (2019): 833-852.

[13] D Praveen Kumar, TarachandAmgoth, and Chandra Sekhara Rao Annavarapu. ,"Machine learning algorithms for wireless sensor networks: A survey.", Information Fusion, 49:1–25, 2019.

[14] G. Ahmed, N. M. Khan, Z. Khalid and R. Ramer, "Cluster head selection using decision trees for Wireless Sensor Networks", *IEEE International Conference on Intelligent Sensors Sensor Networks and Information Processing*, 2008.

[15] Kadhim, A. K., "Intra-clustering communication enhancement in WSN by using skillful methodologies", in Journal of Physics Conference Series, 2020, vol. 1530, no. 1. doi: 10.1088/1742-6596/1530/1/012005.

[16] Dan Li, K. D. Wong, Yu Hen Hu and A. M. Sayeed, "Detection, classification, and tracking of targets," in *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 17-29, March 2002, doi: 10.1109/79.985674.

[17] S. Mostafavi and V. Hakami, "A new rank-order clustering algorithm for prolonging the lifetime of wireless sensor networks", International Journal of Communication Systems, 2019.

[18] 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.

[19] Martyna J. (2008)," An Application of LS-SVM Method for Clustering in Wireless Sensor Networks". In: Nguyen N.T., Katarzyniak R. (eds) New Challenges in Applied Intelligence Technologies. Studies in Computational Intelligence, vol 134. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79355-7_37

[20] Y. Li, Y. Wang, and G. He, ''Clustering-based distributed support vector machine in wireless sensor networks,'' J. Inf. Comput. Sci., vol. 9, no. 4, pp. 1083–1096, 2012.

[21] H. Wang, H. Fang, K. A. Espy, D. Peng, and H. Sharif, "A bayesian multilevel modeling approach for data query in wireless sensor networks," in Computational Science–ICCS 2007. Springer, 2007, pp. 859–866.

[22] O. Franc¸ois, S. Ancelet, and G. Guillot, "Bayesian clustering using hidden markov random fields in spatial population genetics" ,Genetics, vol. 174, no. 2, pp. 805–816, 2006.

[23] Z. Jellali, L. N. Atallah and S. Cherif, "Principal Component Analysis based Clustering Approach for WSN with Locally Uniformly Correlated Data", *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, Tangier, Morocco, 2019, pp. 174-179, doi: 10.1109/IWCMC.2019.8766477.

[24] Muniraju, G., Zhang, S., Tepedelenlioglu, C., Banavar, M. K., Spanias, A., Vargas-Rosales, C., &Villalpando-Hernandez, R. (2017)." Location Based Distributed Spectral Clustering for Wireless Sensor Networks". In *2017 Sensor Signal Processing for Defence Conference, SSPD 2017* (Vol. 2017-January, pp. 1-5). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/SSPD.2017.8233241

[25] Bataineh, Asia K., Mohammad Habib Samkari, AbduallaAbdualla and Saad Al-Azzam. ,"K-Means Clustering in WSN with Koheneon SOM and Conscience Function.", *Mathematical Models and Methods in Applied Sciences* 13 (2019): 63.

[26] Sinha, D., Kumari, R. &Tripathi, S.,"Semisupervised Classification Based Clustering Approach in WSN for Forest Fire Detection.", *Wireless PersCommun* **109,** 2561–2605 (2019). https://doi.org/10.1007/s11277-019-06697-0

[27] A. 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

[28] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction.", The MIT Press, March 1998.

[29] Sumithra, S., Victoire, T.: "Differential evolution algorithm with diversified vicinity operator for optimal routing and clustering of energy efficient wireless sensor networks.", Sci. World J. (2015)