A Hybrid Q-Learning Automata routing protocol for Wireless Sensor Networks

Jakub Gąsior

Department of Mathematics and Natural Sciences
Cardinal Stefan Wyszyński University Warsaw Poland
j.gasior@uksw.edu.pl

Abstract. This paper proposes a novel hybrid approach to routing in Wireless Sensor Networks that combines Q-learning and Learning Automata models. It is designed to optimize the routing process by leveraging the strengths of both techniques: Q-learning's ability to adapt to dynamic network conditions and Learning Automata's fast adaptation and convergence in stable scenarios. Preliminary analysis indicates feasibility of the proposed approach, showing that it can improve the network lifetime and packet delivery ratio when compared with similar routing protocols.

Keywords: Wireless sensor networks · Routing · Learning automata · Q-learning

1 Introduction

Wireless Sensor Network (WSN) is a distributed network comprising small, battery-powered devices called sensors, capable of sensing and collecting data from their surrounding environment. The sensors are typically low-power and have limited computing capabilities, making energy efficiency a crucial aspect of their design. These sensors communicate with one another wirelessly using radio frequency waves and collaborate to perform specific tasks, such as monitoring environmental parameters like temperature, humidity, or air quality. Data is then collected and sent for further processing via a specialized sink node.

Our work will focus on efficient routing methods, allowing for the optimization of the route taken by data from the source to the sink. To that end, we present a hybrid routing model where Q-learning and Learning Automata (LA) are integrated to leverage their respective strengths while minimizing their weaknesses. This combination allows the algorithm to adapt dynamically to the varying conditions of WSNs, ensuring better performance in terms of latency, packet delivery, and routing stability.

The rest of this paper is organized as follows. Section 2 presents the works related to reinforcement learning-based routing algorithms in WSNs. We introduce the theoretical background of the problem in Section 3. Section 4 describes our proposed hybrid Q-LA routing protocol. We present the findings of our experiments in Section 5. The last section concludes the paper with future research directions.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_51

2 Related work

There has been a growing interest in developing reinforcement learning and automata models to tackle the challenges of energy efficiency in WSNs. For example, Manju and Kumar [7] introduced a scheduling algorithm utilizing learning automata to address the target coverage problem. This approach allows sensor nodes to select their operational state autonomously. To validate the efficacy of their scheduling method, comprehensive simulations were conducted, comparing its performance against existing algorithms.

In another study, Lin et al. [5] presented a novel on-demand coverage-based self-deployment algorithm tailored for significant data perception in mobile sensing networks. The authors employed the cellular automata model to accommodate the characteristics of mobile sensing nodes and spatial-temporal node evolution. Subsequently, leveraging learning automata theory and historical node movement data, they proposed a new mobile cellular learning automata model to intelligently and adaptively determine optimal movement directions with minimal energy consumption.

Gudla and Kuda [3] utilized a LA-based model as a routing mechanism for enhanced energy efficiency and reliable data delivery. The approach aims to calculate the selection probability of the next node in a routing path based on various factors such as node score, link quality, and previous selection probability. Furthermore, they proposed an energy-efficient and reliable routing mechanism by combining learning automata with the A-star search algorithm.

Another contribution by Upreti et al. [12] introduced a scheduling technique named Pursuit-LA. Each sensor node in the network was equipped with an LA agent to autonomously determine its operational state to achieve comprehensive target coverage at minimal energy cost. Lastly, Qarehkhani et al. [10] proposed a continuous learning automata-based approach for optimizing sensor angles in Distributed Sensor Networks (DSNs). The method involved continuously adapting sensing angles using LA models. Comparative analysis against a conventional automata-based approach demonstrated the efficacy of the proposed algorithm.

Q-learning, a model-free reinforcement learning algorithm, was also the subject of recent studies in efficient data routing in WSNs. Maivizhi and Yogesh [6] employed it to design a routing algorithm for in-network aggregation (RINA) to build a routing tree based on minimal information such as residual energy, the distance between nodes, and link strength.

Gao et al. [1] employed a Q-learning-based routing optimization algorithm for underwater wireless sensor networks. The authors proposed two reward functions based on the average residual energy of the network, transmission delay, and link success rate to balance transmission quality and lifetime better. A similar solution was proposed by Nandyala et al. [8]. The authors employed the QTAR protocol to determine the next-forwarder candidates along the routing path and adopt Q-learning to aid in the optimal global decision-making of next-hop candidates. It showed a lower energy consumption, shorter latency, and longer network lifetime than other state-of-the-art solutions.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_51

Finally, Jain et al. [4] assessed Q-Learning-based routing based on energy depletion rate, node duration, and packet delivery ratio. The authors proved that reinforcement learning schemes can outperform traditional algorithms such as LEACH and K-means by adopting better energy utilization, reduced node mortality, and higher network throughput.

3 Theoretical background

We consider a WSN comprising N sensors $S = \{s_1, s_2, ..., s_N\}$ randomly deployed over a two-dimensional rectangular area of $x \times y$ $[m^2]$. The area contains M targets $T = \{t_1, t_2, ..., t_M\}$ (also called Points of Interest (POI)) that are uniformly distributed with a step of g. All sensors are assumed to have the same sensing range R_s^i and battery capacity b_i. A Boolean disk represents the coverage model of a sensor node [13] and assumes omnidirectional sensing with no random variations.

We use a bipartite graph $G = (V, E)$ to model the Target Coverage Problem (TCP), with $V = S \cup T$, where S represents a set of sensor nodes, T a set of targets and E the set of edges as follows: $\{s, t\} \in E$ if and only if the sensor node s_i detects the target t_j. We define the degree $d(t_j)$ of the target t as the number of sensor nodes that detect the target t_j.

Further, we employ the first-order radio model for the sensors and assume the energy spent for transmitting and receiving a data packet is constant. In addition to this, the energy expenditure is proportional to the distance between two nodes [6].

3.1 Q-learning

Q-learning is a model-free reinforcement learning algorithm used for sequential decision-making, especially when the environment is uncertain or dynamic. It learns optimal policies over time by adjusting its action-value function based on rewards for selecting specific actions (in our case, selecting next-hop nodes). The core components of the Q-learning algorithm can be defined as follows [8]:

1. Define the State Space ($S = \langle E, L, C, D \rangle$) by assigning the variables that affect routing in the WSN, i.e.:
   - Energy level (E): remaining energy of the node,
   - Link quality (L): measured by the Packet Delivery Ratio (PDR),
   - Congestion level (C): the number of packets currently in the node's buffer or the average delay,
   - Distance to destination (D): distance to the sink or destination node.
2. Define the Action Space (A) corresponding to the set of potential decisions a node can make, i.e., the next-hop nodes;
3. Initialize the Q-values for each state-action pair to an arbitrary value (e.g., zero);
4. Routing Decision (Action Selection) are selected based on the ϵ-greedy policy:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_51

- Exploration: with probability ϵ, the node randomly selects an action from the set of possible actions,
- Exploitation: with probability $1 - \epsilon$, the node selects the action with the highest Q-value (best-known action based on previous experiences).

5. Once the action is chosen (i.e., the next-hop node is selected), the node forwards the packet and observes the reward based on the outcome of the transmission:
   - A reward is given if the packet is successfully transmitted to the next hop,
   - A penalty is assigned if the transmission fails (due to poor link quality, congestion, or node energy depletion).
6. After observing the outcome, the Q-value for the state-action pair is updated using the Bellman equation [1]:

$$Q^{new}(s_t, a_t) = Q(s_t, a_t) + \alpha[R_t + \gamma \times max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)], \quad (1)$$

   where:
   - α is the learning rate;
   - γ is the discount factor;
   - R_t is the reward observed after performing action a_t;
   - $max_{a'} Q(s_{t+1}, a')$ is the maximum Q-value of the next state s_{t+1}, which represents the best possible future reward.

   This update process gradually refines the Q-values, leading to better routing decisions.

This solution provides several advantages, such as adaptability to dynamic network conditions (e.g., node mobility, varying link quality, energy constraints) by continually learning the optimal routing paths. Due to its decentralized nature, each node can independently learn from its environment without requiring global knowledge, making it scalable and suitable for larger networks [1, 4, 11].

Some challenges need to be acknowledged. One of the main problems is balancing exploration and exploitation (controlled by the parameter ϵ). Too much exploration can lead to inefficient routing, while too much exploitation can cause the network to get stuck in suboptimal paths. Additionally, Q-learning requires maintaining and updating a Q-value table, which can be computationally expensive, especially in large-scale networks. Thus, converging to the optimal policy may take a long time, especially in highly dynamic environments.

3.2 Learning automata

A learning automaton is a self-operating mechanism that responds to a sequence of instructions in a certain way to achieve a particular goal. The automaton either responds to a predetermined set of rules or adapts to the environmental dynamics in which it operates [9]. We define the environment influencing the activities of the automaton as a triple $E = \langle A, C, B \rangle$, where:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_51

– $A = \alpha_1, \alpha_2, \ldots, \alpha_r$ is the set of actions;
– $B = \beta_1, \beta_2, \ldots, \beta_m$ is the output set of the environment. When $m = 2$, $\beta = 0$ corresponds a reward and $\beta = 1$ represents a penalty;
– $C = c_1, c_2, \ldots, c_r$ is a set of punishment or penalty probabilities, where $c_i \in C$ corresponds to an input activity α_i.

The learning process involving the LA and a random environment is presented in Fig. 1. Whenever an automaton generates an action α_t, the environment sends a response β_t either penalizing or rewarding the automaton with a specific probability c_i.

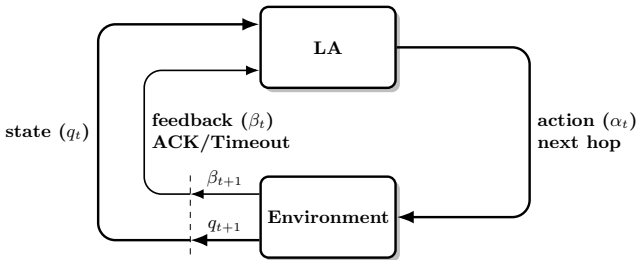

Fig. 1: A feedback loop of learning automata.

Generally, LA can be categorized as a fixed structure LA or a variable structure LA. This paper considers variable structure LA, where the action probability vector is not fixed, and the action probabilities are updated after each iteration. Thus, through interactions with the environment, LAs may adjust their action-selection probabilities by a positive reinforcement (i.e., Reward, Eq. (2)):

$$\begin{aligned} p_i(t+1) &= p_i(t) + a(1 - p_i(t)) && j = i \\ p_j(t+1) &= (1-a)p_j(t) && \forall j, j \neq i \end{aligned} \tag{2}$$

or a negative reinforcement (i.e., Penalty, Eq. (3)):

$$\begin{aligned} p_i(t+1) &= (1-b)p_i(t) && j = i \\ p_j(t+1) &= \frac{b}{r-1} + (1-b)p_j(t) && \forall j, j \neq i \end{aligned} \tag{3}$$

Values $p_i(t)$ and $p_j(t)$ are the probabilities of actions α_i and α_j at time t, r is the number of actions, while a and b are the reward and the penalty parameters, respectively. We employ a learning algorithm called *Linear Reward-Penalty* (L_{R-P}) with $a = b$ in our work [9].

4 A Hybrid Q-LA routing protocol

In order to mitigate the inherent challenges present within the Q-Learning algorithm, we propose the hybrid approach, which dynamically combines Q-Learning and Learning Automata to make routing decisions based on network

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_51

conditions such as energy levels, link quality, and congestion. The idea is to use Q-Learning for adaptive, long-term decision-making (in high-congestion situations) and Learning Automata for short-term adjustments when the network is stable.

The hybrid model leverages Q-Learning when the network is highly dynamic or when energy depletion is significant (for better adaptation). On the other hand, Learning Automata is favored when the network is stable, as it allows quicker convergence with fewer computational overheads [2].

At each decision-making step, nodes switch between Q-Learning and Learning Automata based on network conditions. If the node runs low energy, it will use Q-learning to optimize routing decisions and conserve power. Similarly, if the queue length is high (overloaded network), Q-learning can help find alternative, less congested routes. On the other hand, LA is more efficient when the network does not experience frequent link failures or congestion. It adjusts probabilities of choosing the next hop based on local conditions without needing extensive state exploration.

Every sensor node s_i maintains a Q-table where each entry represents each neighbor's expected reward (path quality). The algorithm then adjusts the Q-values to improve the routing decisions over time. The action with the highest Q-value will be selected (i.e., next-hop with the highest potential for success) with probability $1 - \epsilon$. Additionally, nodes maintain a probability distribution when selecting each neighboring node (Eqs. (2) and (3)). These probabilities are dynamically updated based on reward signals from the environment, where successful packet delivery increases the probability of selecting a particular neighbor, and packet loss or failed delivery decreases the probability.

When conditions are stable, the node uses Learning Automata for quicker decision-making and convergence. When unstable conditions occur (e.g., low energy or congestion), it switches to Q-learning for optimal path discovery.

5 Experimental Study

In this section, we aim to evaluate the effectiveness of the proposed hybrid Q-LA routing protocol through multiple computer simulations. To accomplish this, we will employ a custom WSN simulator written in Matlab. We use a fixed network, where sensor nodes are randomly positioned within a $1000 : m \times 1000 : m$ area alongside a static deployment of $T = 400$ targets. The sensing range of sensors was set at a value of $R_s^i = 175$. The number of nodes will vary in the range $S = \{100, 150, 200, 250, 300\}$ sensors.

We will compare its performance with multiple routing protocols, including LEACH (Low Energy Adaptive Clustering Hierarchy), AODV (Ad hoc On-demand Distance Vector), and RPL (Routing Protocol for Low Power and Lossy Networks). The performance of the basic Q-Learning routing scheme (without LA improvements) will serve as a benchmark against the proposed solution. We will be evaluating the routing efficiency based on performance metrics listed below:

– Packet Delivery Ratio (PDR): the ratio of packets delivered successfully to the destination node (sink) over the total packets sent;

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_51

– Latency: the average time a packet takes from the source node to the destination node (sink).
– Energy Consumption: the total energy consumed by the network, considering both the transmission and reception of packets;
– Network Lifetime: the duration for which the network remains operational before the first node runs out of energy.

The experiment results are presented in Table 1. There were averaged over 30 runs to ensure robustness. Through this evaluation, we seek insights into the algorithms performance across various network conditions.

Table 1: Averaged results of comparative performance metrics for tested routing protocols.

Protocol	PDR [%]	Latency [ms]	Energy [%]	Lifetime [t]
Q-Learning	86.23	74.51	**21.64**	63
Hybrid Q-LA	**89.42**	**65.59**	23.66	**68**
AODV	79.23	**75.34**	**24.75**	59
RPL	**83.58**	88.71	29.46	**61**
LEACH	75.43	91.47	32.77	54

As stated before, the hybrid model dynamically selects the most appropriate routing technique based on the network's current state. Q-Learning provides long-term adaptability to changing network conditions (i.e., failures, congestion), while Learning Automata ensures faster, local adjustments during stable conditions.

By dynamically switching between these two approaches, the algorithm provides a higher delivery success rate, lower latency, and longer network lifetime at the cost of a slight increase in overall energy consumption. Regardless of the variant, the reinforcement learning-based protocols offer better overall efficiency than the standard routing solutions.

6 Conclusion

This paper presents a novel Q-Learning and Learning Automata (Q-LA) routing protocol for Wireless Sensor Networks. Our early research findings demonstrate that this hybrid Q-LA approach provides a robust and adaptive solution to wireless sensor networks' dynamic and unpredictable nature. Learning optimal routing policies based on the local conditions of each node improves packet delivery, latency, and network lifetime. Though there are challenges related to convergence and computational overhead, proper parameter tuning could significantly enhance the performance of routing protocols in WSNs. Our future work will include

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_51

further testing in real-world WSNs, especially in interference-prone environments, by introducing link failures and node mobility.

Bibliography

[1] J. Gao, J. Wang, et al. "Q-Learning-Based Routing Optimization Algorithm for Underwater Sensor Networks". In: *IEEE Internet of Things Journal* 11.22 (2024), pp. 36350–36357. DOI: 10.1109/JIOT.2024.3398797.

[2] J. Gąsior. "Learning Automata Strategies for Prolonging Lifetime of Wireless Sensor Networks". In: *Advances in Practical Applications of Agents, Multi-Agent Systems, and Digital Twins: The PAAMS Collection*. Ed. by P. Mathieu and F. De la Prieta. Springer Nature Switzerland, 2025, pp. 109–120. ISBN: 978-3-031-70415-4.

[3] S. Gudla and N. R. Kuda. "Learning automata based energy efficient and reliable data delivery routing mechanism in wireless sensor networks". In: *Journal of King Saud University - Computer and Information Sciences* 34.8, Part B (2022), pp. 5759–5765. ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2021.04.006.

[4] A. Jain, S. Jain, and G. Mathur. "Optimizing wireless sensor network routing with Q-learning: enhancing energy efficiency and network longevity". In: *Engineering Research Express* 6 (Nov. 2024). DOI: 10.1088/2631-8695/ad9138.

[5] Y. Lin, X. Wang, et al. "An on-demand coverage based self-deployment algorithm for big data perception in mobile sensing networks". In: *Future Generation Computer Systems* 82 (2018), pp. 220–234. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2018.01.007.

[6] R. Maivizhi and P. Yogesh. "Q-learning based routing for in-network aggregation in wireless sensor networks". In: *Wirel. Netw.* 27.3 (Apr. 2021), pp. 2231–2250. ISSN: 1022-0038. DOI: 10.1007/s11276-021-02564-8.

[7] S. Manju and B. Kumar. "Target coverage heuristic based on learning automata in wireless sensor networks". In: *IET Wireless Sensor Systems* 8.3 (2018), pp. 109–115. DOI: https://doi.org/10.1049/iet-wss.2017.0090.

[8] C. S. Nandyala, H.-W. Kim, and H.-S. Cho. "QTAR: A Q-learning-based topology-aware routing protocol for underwater wireless sensor networks". In: *Computer Networks* 222 (2023), p. 109562. ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2023.109562.

[9] K. S. Narendra and M. A. L. Thathachar. *Learning automata: an introduction*. USA: Prentice-Hall, Inc., 1989. ISBN: 0134855582.

[10] A. Qarehkhani, M. Golsorkhtabaramiri, et al. "Solving the target coverage problem in multilevel wireless networks capable of adjusting the sensing angle using continuous learning automata". In: *IET Communications* 16.2 (2022), pp. 151–163. DOI: https://doi.org/10.1049/cmu2.12323.

[11] V. K. Sharma, S. S. P. Shukla, and V. Singh. "A tailored Q- Learning for routing in wireless sensor networks". In: *2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing*. 2012, pp. 663–668. DOI: 10.1109/PDGC.2012.6449899.

[12] R. Upreti, A. Rauniyar, et al. "Adaptive pursuit learning for energy-efficient target coverage in wireless sensor networks". In: *Concurrency and Computation: Practice and Experience* 34.7 (2022). DOI: https://doi.org/10.1002/cpe.5975.

[13] B. Wang. "Coverage Problems in Sensor Networks: A Survey". In: *ACM Comput. Surv.* 43.4 (Oct. 2011). ISSN: 0360-0300. DOI: 10.1145/1978802.1978811.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_51