Research Article

Mathematical Approach Of Q-Learning With Temporal Difference Method In Sensor Data Communication In Cloud Environment

[1]P. Abirami, [2]Dr. S. Vijay Bhanu, [3]Dr.T.K.Thivakaran

[1]Research Scholar
Department of Computer Science and Engineering
Annamalai University
abiramipadmanaban.research@gmail.com
[2]Associate Professor and Research Supervisor
Department of Computer Science and Engineering
Annamalai University
svbhanu22@gmail.com
[3]Professor
Department of Information Technology
S.R.M University
Chennai
tktcse4@gmail.com
*corresponding author: *abiramipadmanaban.research@gmail.com*

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Access to data is efficient and can be managed with minimal sensors. However the sensor data captured at different locations can be integrated to the cloud[11]In this work, we derived the optimum path, which is the shortest, and the time and power are optimized. We consider the network topologies where nodes and paths are assigned by a definite probability. The Bellman equation and the temporal difference method are used in Q-learning. Bellman -Temporal based algorithm is used for finding the optimal path among users in the cloud environment.

Keywords: Wireless body sensor networks, Q-Learning algorithm, Q-function, Temporal difference Learning,and Bellman equation.

I. INTRODUCTION

A sensor network has a set of sensor nodes. The sensor nodes collect information and is processed for analysis. The external sources also present data that can be gathered for analysis. The data gathered from health care systems are important for analyzing the patient progress. The data is secured in the cloud environment[10][12]. A possibility of finding the optimal path will help in improving the efficiency of the trust among cloud users[11].

"Q-Learning is a value-based reinforcement learning algorithm which is used to find the optimal action-selection policy using a Q function. Weaim is to optimize the value function Q. The Q matrix helps us to find the best action for each state.The Q-function $Q: S \times A \to \Re$ uses the Bellman equation and takes two inputs state and action $Q(s_t, \alpha_t) = E(R_{t+1} + \lambda R_{t+2} + \lambda^2 R_{t+3} + ... \mid s_t, \alpha_t)$".

The proposed algorithm has a map that evaluates the best of a "state-action combination".

First, initialize Q with an arbitrary fixed value. Then, at each time t the agent selects an action α_t ,observes a reward R_w, enters a new state s_{t+1} ("that may depend on both the previous state s_t and the selected action"), and Q is updated. The center of the algorithm is a Bellman equation as a simple value iteration update, using the average of the old value and the new value. The algorithm ends when state s_{t+1} is a final or terminal state[1]."Temporal difference learning refers to a class of model free reinforcement learning methods which learn by bootstrapping from the current estimate of the value function. These methods sample from the environment and perform updates based on current estimates. Temporal difference methods adjust predictions to match later, more accurate, predictions about the future before the final outcome is known".

II. THE BELLMAN'S EQUATION

The Bellman equation is a functional equation,it finds the value function *V*. The value function describes the best possible value of the objective, as a function of the state s. From the value function, we will also find the function $\alpha(s)$ that describes the optimal action as a function of the state.

The Bellman equation can be described as a recursive function:

$$V(s_t) = \underset{\alpha}{Max}\{R_w(s_t,\alpha) + \lambda V(s_{t+1})\}, \text{ where,}$$

S_t : Current state,

α : Action,

S_{t+1} : Next state,

λ : Discount factor,

$R_w(s_t,\alpha)$: Reward function,

$V(s)$: Valueof acurrent state.

If the chance of moving from the state S_t to the state S_{t+1} with action α is $P(s_t,\alpha,s_{t+1})$ then the bellman equation becomes

$$V(s_t) = \underset{\alpha}{Max}\left\{R_w(s_t,\alpha_t) + \lambda\sum_{s'} P(s_t,\alpha,s_{t+1})V(s_{t+1})\right\}$$

Therefore the Q-function is

$$\Rightarrow Q(s_t,\alpha_t) = \underset{\alpha}{Max}\left\{R_w(s_t,\alpha_t) + \lambda\sum_{s'} P(s_t,\alpha,s_{t+1})V(s_{t+1})\right\}$$

$\sum_{s'} P(s_t,\alpha,s_{t+1})V(s_{t+1})$ is the mean value,

$$\Rightarrow Q(s_t,\alpha_t) = R_w(s_t,\alpha_t) + \lambda\sum_{s'} P(s_t,\alpha,s_{t+1})\,\underset{\alpha}{Max}\,Q(s_{t+1},\alpha_{t+1})$$

$$\Rightarrow Q(s_t,\alpha_t) = R_w(s_t,\alpha_t) + \lambda\,\underset{s'}{Max}\,Q(s_{t+1},\alpha_{t+1})$$

Now, the temporal difference:

$$T_d = R_w(s_t,\alpha_t) + \lambda\,\underset{s'}{Max}\,Q(s_{t+1},\alpha) - Q(s_t,\alpha_t)$$

The optimum Q value,

$$Q_n(s_t,\alpha_t) = Q_o(s_t,\alpha_t) + \beta * T_d(s_t,\alpha_t)$$

III. MATHEMATICAL MODEL TO THE PROBLEM

Let Z_i *and* X_{ji}, $i = 1,2,\ldots,n$ be the zones and the cluster-heads in these zones respectively. Let $\alpha_i, i = 1,2,\ldots,p$ be the set of routes present in every node. When a signal is in an exacting node there is an option of choosing any one route from the "p" probable routes with large chance that are nearby[10].

When a signal reaches a node it can obtain a reward for reaching that node and is called as the immediate reward $R_w(X_{ji},\alpha_j)$, which is got by selecting the route α_j. ("The node is given an option of choosing the route, in such a way that the packet will get a maximized cumulative reward, which the packet can gain when it moves from the current location node to a new node, thereby trying to reach the destination in the shortest path".)[2][3]

If the signal is in node X_{ji} and if it chooses its next node positioned at X_{jj} then the Q value at the present node, $X_{jj,}$ can be obtained from the following relation.

$$Q_n(s,\alpha) = Q_o(s,\alpha) + \beta * T_d(s,\alpha);$$

$$T_d(s,\alpha) = R_w(s,\alpha) + \lambda * Max\{Q(\mathrm{N_s},\gamma)\} - Q(s,\alpha)$$

Where,
$S : Currentstate$; $\alpha : Action$; $\lambda : discountfactor$,
$R_w(s,\alpha) : Reward$,
$N_s : Nextstate$;
$\gamma : allpossibleAction$;

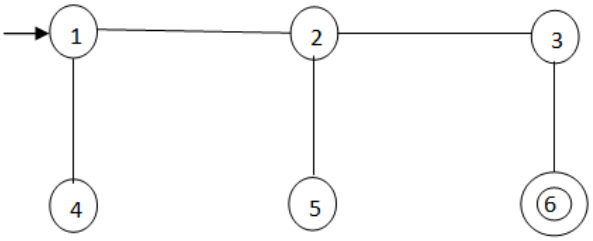
$T_d(s,\alpha) : Temporal\ distance$;
$\beta : Learningrate$;
$\mathrm{Q_n}(s,\alpha)$: $NewQ\_value$;
$\mathrm{Q_0}(s,\alpha)$: $OldQ\_value$.
Note: Q Parameter is used to find an optimal path for the packets in a wireless body sensor network[4][9]:

Proposed Algorithm:

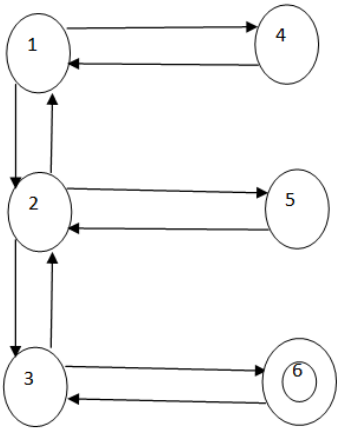
1. Take the reward matrix R.
2. Initialize the Q matrix as a null matrix.
3. Take anaction from Q table for the initial state.
4. Perform the chosen action and transition to the next state.
5. Get the reward
6. Compute the Temporal difference from
$$T_d(s,\alpha) = Rw(s,\alpha) + \lambda * Max\{Q(Ns,\gamma)\} - Q(s,\alpha)$$
7. Evaluate : $Q_n(s,\alpha) = Q_o(s,\alpha) + \beta * T_d(s,\alpha)$
8. Replicate all the steps from step 3 till the current state and final state same.
9. Stop.

Mathematical Example:

Consider the Zone Z_i which has four nodes and they are linked as follows:



The corresponding state diagram is given below, and take the state-1 is the current state and the state-6 is the goal state[5].



The reward Matrix R_w of the given state diagram:

$$R = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Select the initial state as signal node 1, and the initial Q matrix is taken as a zero matrix[6].

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The row and column of the Q matrix represents the present state and the probable action foremost to the next state.

Now, the temporal distance and the new Q-value are calculated from the following equations:

$$T_d(s,\alpha) = Rw(s,\alpha) + \lambda * Max\{Q(Ns,\gamma)\} - Q(s,\alpha)$$

$$Q_n(s,\alpha) = Q_o(s,\alpha) + \beta * T_d(s,\alpha)$$

Take the discount factor and the learning rate as $\lambda = 0.9,\ \beta = 0.8;$

and $\lambda, \beta \in [0,1]$, when $\lambda \approx 0$, will consider only immediate reward, and when $\lambda \approx 1$ will consider the future rewards with greater weight[7].

Now, the current state = initial state = 1.

(s,α)	$T_d(s,\alpha)$	$Q_n(s,\alpha)$	Q-Matrix
(1,2)	1	0.8	$\begin{pmatrix} 0 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

(1,4)	1.72	1.376	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
(2,1)	1	0.8	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
(2,3)	1.72	1.376	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 1.376 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
(2,5)	2.2384	1.79072	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 1.376 & 0 & 1.79072 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
(3,2)	1	0.8	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 1.376 & 0 & 1.79072 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
(3,6)	1.72	1.376	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 1.376 & 0 & 1.79072 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 1.376 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

(4,1)	1	0.8	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 1.376 & 0 & 1.79072 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 1.376 \\ 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
(5,2)	1	0.8	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 1.376 & 0 & 1.79072 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 1.376 \\ 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
(6,3)	1	0.8	$\begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 1.376 & 0 & 1.79072 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 1.376 \\ 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0 \end{pmatrix}$

Now,

$$Q = \begin{pmatrix} 0 & 0.8 & 0 & 1.376 & 0 & 0 \\ 0.8 & 0 & 1.376 & 0 & 1.79072 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 1.376 \\ 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0 \end{pmatrix}$$

Hence,the highest Q value is 1.79072.The optimal path is 2-5.

IV. CONCLUSION

The paper has proposed a efficient path based data transmission. The model proposed is efficient as it captures the path over time using Q-learning. The data captured can be used and managed in a optimal manner[8]. From the solution above proposed with Bellman -Temporal based algorithm ,the optimal path is reached with a higher Q-Score. This can be extended in future as a complete security solution with Q-learning model for the cloud environment.

REFERENCES

1. M. Blount, V. Batra, A. Capella, M. Ebling, W. Jerome, S. Martin, M. Nidd, M. Niemi and S. P. Wright, “Remote Healthcare Monitoring using Personal Care Connect”, IBM System Journal, Vol. 46, No. 1, pp. 95-113, 2007.
2. S. Ivanov, C. Foley, S. Balasubramaniam and D. Botvich, “Virtual Groups for Patient Wireless Body Area Network Monitoring in Medical Environments”, IEEE Transactions on Biometric Engineering, Vol. 59, No. 11, pp. 3238-3246, 2012.
3. B. latre, B. Braem, I.Moeman, “A survey on wireless body area networks”, Wireless Networks, Springer, 17(1), pp. 1-18, 2011.
4. Miroslav Chleb´ık and JankaChleb´ıkov´a. Approximation hardness of edge dominating set problems. Journal of Combinatorial Optimization, 11(3):279– 290, 2006. [30] Va˘sekChv´at

5. Raghavendra C S, Sivalingam K M, Znati T. Wireless Sensor Networks. Dordrecht: Kluwer Academic Publishers, 2004.
6. Ravi R. Rapid rumor ramification: approximating the minimum broadcast time. In: Proceedings of the 35-th IEEE Annual Symposium on Foundations of Computer Science. 1994, 202–213.
7. A. Roy and K. Das, "QM2RP: A QoS-based Mobile Multicast Routing Protocol using Multi-Objective Genetic Algorithm", Wireless Networks, Vol. 10, No. 3, pp. 271-286, 2004.
8. S. Sara, S. Prasanna and D. Sridharan, "A Genetic Algorithm based Optimized Clustering for Energy-Efficient Routing in MWSN", ETRI Journal, Vol. 34, No. 6, pp. 922-931, 2012
9. Xu L, Xiang Y, Shi M. On the problem of channel assignment for multi-NIC multiple wireless networks. Lecture Notes in Computer Sciences, 2005, 3794: 633– 642.
10. Zhu J, Chen X, Hu X. Minimum multicast time problem in wireless sensor networks. Lecture Nodes in Computer Sciences, 2006, 4138: 490–501.
11. Abirami, P., Bhanu, S.V. Enhancing cloud security using crypto-deep neural network for privacy preservation in trusted environment. Soft Computing , 24, 18927–18936 (2020). https://doi.org/10.1007/s00500-020-05122-0
12. Dash S.K., Sahoo J.P., Mohapatra S., Pati S.P. (2012) Sensor-Cloud: Assimilation of Wireless Sensor Network and the Cloud. In: Meghanathan N., Chaki N., Nagamalai D. (eds) Advances in Computer Science and Information Technology. Networks and Communications. CCSIT 2012. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 84. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27299-8_48