BILINGUAL PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Data Analytics of an Information System Based on a Markov Decision Process and a Partially Observable Markov Decision Process

Lidong Wang**, Reed L. Mosher, Terril C. Falls, Patti Duett

Institute for Systems Engineering Research, Mississippi State University, Vicksburg, MS 39180, USA

ABSTRACT

Data analytics of an information system is conducted based on a Markov decision process (MDP) and a partially observable Markov decision process (POMDP) in this paper. Data analytics over a finite planning horizon and an infinite planning horizon for a discounted MDP is performed, respectively. Value iteration (VI), policy iteration (PI), and Q-learning are utilized in the data analytics for a discounted MDP over an infinite planning horizon to evaluate the validity of the MDP model. The optimal policy to minimize the total expected cost of states of the information system is obtained based on the MDP. In the analytics for a discounted POMDP over an infinite planning horizon of the information system, the effects of various parameters on the total expected cost of the information system are studied.

Keywords: Predictive modelling; Information system; MDP; POMDP; Cybersecurity; Q-learning

1. Introduction

Cyberattacks against federal information systems in the USA are more and more sophisticated. The probability of grave damages keeps increasing in spite of efforts and the use of substantial resources. There are challenges in completely aggregating heterogeneous data from various security tools, analyzing the collected data, prioritizing remediation activities, and reporting in an approach to directing a suitable response [1]. Cyberspace is a dynamic environment. Targets are not always static. No offensive or defensive capability keeps being indefinitely effective. There is no permanent advantage [2].

Cyber attackers generally have advantages over the defender of an information system. The advantages lie in: 1) Attackers can choose the place and time of an attack; 2) Attackers can only exploit a sin-

*CORRESPONDING AUTHOR:
Lidong Wang, Institute for Systems Engineering Research, Mississippi State University, Vicksburg, MS 39180, USA; Email: lidong@iser.msstate.edu

ARTICLE INFO
Received: 26 January 2023 | Revised: 17 February 2023 | Accepted: 20 February 2023 | Published Online: 28 February 2023
DOI: https://doi.org/10.30564/jcsr.v5i1.5434

CITATION
Wang, L.D., Mosher, R.L., Falls, T.C., et al., 2023. Data Analytics of an Information System Based on a Markov Decision Process and a Partially Observable Markov Decision Process. Journal of Computer Science Research. 5(1): 21-30. DOI: https://doi.org/10.30564/jcsr.v5i1.5434

COPYRIGHT
Copyright © 2023 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License. (https://creativecommons.org/licenses/by-nc/4.0/).

gle vulnerability while the defender has a much more costly task of mitigating all kinds of vulnerabilities. Human-centered cyber-defense practices have not kept pace with threats of targeting and attacking organizations. An integrated approach is needed to speed up detection or responses and slow down attacks. Security automation and intelligence sharing can reduce the defender's costs and save time. Information sharing helps improve the efficiency in detecting and responding to cyberattacks [3].

There are four major categories of attacks [4-6]: 1) Denial of service—trying to stop legitimate users from utilizing services; 2) Probe—trying to get the information of a target host; 3) User to Root (U2R)—unauthorized access to privileges of a local super-user (root); and 4) Remote to Local (R2L)—unauthorized access from a remote machine. Signature-based detection and anomaly-based detection are the two main methods of detecting attacks. Signature-based detection uses predefined attack specifications that are clear and distinct signatures. The database of signatures needs to be updated when there are new signatures. Human security experts are generally required to analyze data related to attacks manually and formulate specifications regarding attacks [7]. Anomaly-based detection is also called behavior-based detection. It models behaviors of the network, computer systems, and users; and raises an alarm when there is a deviation from normal behaviors [8].

Many cyberattacks are characterized by a high level of sophistication. Typically, an advanced persistent threat (APT) is a kind of attack targeting an asset or a physical system with high values. APT attackers frequently leverage stolen credentials of users or zero-day exploits to avoid triggering alerts. This kind of attacks could continue over an extended period of time [9]. Artificial intelligence (AI) or intelligent agents are needed to fight attack, especially an APT. Therefore, the mechanisms of cyber defense should be 1) increasingly intelligent, 2) very flexible, and 3) robust enough to detect various threats and mitigate them. Much research has been done on intrusion detection and prevention systems. Various methods and algorithms of artificial intelligence have been used for cybersecurity. The algorithms include support vector machines (SVM), convolution neural networks, recursive neural networks, general artificial neural networks (ANN), Q-learning (QL), decision trees (DT), *k*-means, *k*-nearest neighbors (*k*-NN), etc. [10]. MDP and POMDP are used in this paper because they deal with the optimal policy or actions based on computed benefits or costs.

During an attack, both the attacker and the defender are in the process of learning about each other. The knowledge evolution of the attacker and the defender indicates the process of learning. A defender's knowledge includes, for example, attackers' objectives, methods utilized, possible technical levels, etc. An attacker's knowledge can be the topology of a defender's network or information system, the operating system version and applications running on servers, etc. When an attack is detected, the defender can expel the attacker or keep it in the information system in order to observe or learn about it. The policy of always expelling the attacker is not optimal in many situations. There is a trade-off between the opportunity of learning about the attacker and the risk of the attacker's damage during the defender's learning process [11]. MDP and POMDP can handle the trade-off and decide on optimal policies or actions.

This paper aims to conduct analytics of an information system based on an MDP and a POMDP. Various methods and algorithms were used, including value iteration (VI), policy iteration (PI), and Q-learning in the analytics of a discounted MDP over an infinite planning horizon to evaluate the MDP model validity and parameters in the model. In the modelling of a discounted POMDP over an infinite planning horizon, the effects of several important parameters on the total expected reward of the system were studied. The data analytics of the MDP and POMDP in this paper was conducted using the *R* language and its functions. The organization of this paper is as follows: the next section introduces the methods of MDP; Section 3 introduces the methods of POMDP; Section 4 presents an MDP model of an information system; Section 5 shows the analytics of the information system based on MDP; Section 6 presents the analytics of the information system based on POMDP; and the final section is the conclusions.

2. Markov decision process

An MDP can be defined by a tuple $<S, A, P, R, \gamma>$ [12-14]: S refers to a set of states; A is a set of actions; P represents a transition probability matrix that describes the transition from state s to state s' ($s \in S, s' \in S$) after action a ($a \in A$); R refers to the immediate reward after action a; and γ ($0 < \gamma < 1$) is a discounted reward factor. Solving an MDP is often a process of finding an optimal policy to maximize the total expected reward or minimize the total expected cost.

Policy iteration, value iteration, and Q-learning are often used to obtain an optimal policy for an MDP. Data analytics results based on the algorithms of the three methods may be noticeably different, or there can be convergence problems during iterations if the MDP model is not reasonable due to unsuitable model parameters or an incorrect model structure. Therefore, the three methods are employed in this paper, and results are compared to evaluate the model's validity.

PI tries to find a better policy (compared to the previous policy). An iterative process of policy evaluation and policy improvement is stopped when two successive policy iterations result in the same policy, indicating the optimal policy is achieved. The policy iteration is described in Algorithm 1 [15,16]. $P(s, a, s')$ is the probability of the transition. $R(s, a, s')$ is the immediate transition reward from the state s to the state s' after the action a. $V(s)$ and $V(s')$ are the expected total reward of state s and state s', respectively. $\pi(s)$ is an optimal policy of state s.

An optimal policy of the MDP can also be achieved by utilizing VI [15,17]. The stopping criterion is that the value difference of two successive iterative steps is less than the tolerance τ (a very small positive number). Algorithm 2 shows the value iteration process.

Algorithm 1. Policy Iteration.

1	Initial policy Choose an initial policy arbitrarily for all $s \in S$ $V(s) \in R$ and $\pi(s) \in A$
2	Policy evaluation Repeat $\Delta \leftarrow 0$ For each $s \in S$ $v \leftarrow V(s)$ $V(s) \leftarrow max_a \sum_{s'} P(s, \pi(s), s')(R(s, \pi(s), s') + \gamma V(s'))$ $\Delta \leftarrow \max(\Delta, \lvert V(s) - v \rvert)$ until $\Delta < \tau$ (a very small positive number)
3	Policy improvement routine For each state s $\pi(s) \leftarrow argmax_a(\sum_{s'} P(s, a, s')(R(s, a, s') + \gamma V(s')))$
4	Stopping rule If policy is stable, then stop; else go to step 2

Algorithm 2. Value Iteration.

1	Initialization Select $V(s)$ arbitrarily (e.g., $V(s) = 0$ for all $s \in S$)
2	Value iteration process Repeat $\Delta \leftarrow 0$ For each $s \in S$ $v \leftarrow V(s)$ $V(s) \leftarrow max_a \sum_{s'} P(s, \pi(s), s')(R(s, \pi(s), s') + \gamma V(s'))$ $\Delta \leftarrow \max(\Delta, \lvert V(s) - v \rvert)$ until $\Delta < \tau$
3	Output the optimal policy and the maximal values of $V(s)$

Q-learning [17,18] enables an agent to learn the Q-value function which is an optimal action-value function. It can be employed to solve a discounted MDP. Specifically, it is used to compute the expected total reward (or cost) and find the optimal policy in this paper. It can be used to perform data analytics and simulation of a discounted MDP over an infinite planning horizon if the number of iterations to perform is large enough. A Q-learning algorithm is shown in Algorithm 3. $Q(s,a)$ is the action-value function. $\beta \in (0, 1)$ is the learning rate and it is often chosen to be decreased appropriately, e.g., $\beta = 1/\sqrt{(n+2)}$ (n is the iteration step number or the epoch number). The iterative process and the Q-learning update continue until the final step of an episode. The best action a at state s is chosen according to the optimal policy $\pi(s)$.

Algorithm 3. Q-learning.

1	Initialization Initialize $Q(s,a)$ arbitrarily (e.g., $Q(s,a) = 0$, $\forall s \in S, \forall a \in A$)
2	Iterative process and Q-learning update Repeat For each $s \in S$ $Q(s,a) \leftarrow \sum_{s'} P(s,a,s')(R(s,a,s') + \gamma V(s'))$ Q-learning update is as follows: $Q(s,a) \leftarrow (1-\beta)Q(s,a) + \beta[R(s,a,s') + \gamma \max_a Q(s',a)]$ until the final step of episode
3	Output the optimal policy and maximal values of states

3. Partially observable Markov decision process

In many applications, a POMDP is a more realistic model than the classic MDP [19]. The transition model $P(s'|s, a)$, actions $A(s)$, and the reward function $R(s, a, s')$ in a POMDP are the same elements as those in an MDP. The optimal action of the POMDP depends only on the agent's current belief state. The agent does not know its real state; all it knows is the belief state [20]. Besides the three elements, there are a set of observations $O = \{o_1, o_2, ..., o_k\}$ and a set of conditional observation probabilities $B(o|s', a)$ in a POMDP [21].

If b was the previous belief state, and the agent takes action a and then perceives evidence o, then the new belief state [20] is obtained using the following formula:

$$b'(s') = \alpha P(o|s') \Sigma_s P(s'|s, a)\, b(s) \quad (1)$$

where a is a normalizing constant, making the belief state sum to 1.

The optimal value of any belief state b is the infinite expected sum of discounted rewards starting in state b, and executing the optimal policy. The value function, $V^*(b)$, is expressed as follows [22]:

$$V^*(b) = max_{a \in A}\left[b(s)R(s,a) + \gamma \sum_{o \in O} P(o|b,a)V^*(b')\right] \quad (2)$$

4. A Markov decision process model of the information system

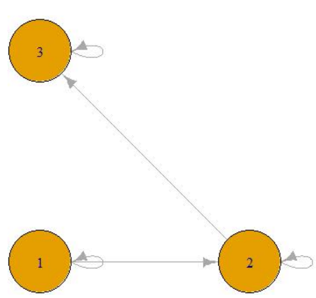
4.1 The structure of the MDP model

The information system has the following states: State 1—no attacker is connected to the information system; state 2—an attacker is connected to the information system, but it has not been detected; and state 3—the attacker is detected. The defender needs to make a decision: wait (no action) or expel only when an attack is detected (state 3). After an expelling action, the system will return to state 1.

The MDP model of the information system is established. State transitions among three states (states 1-3) of two decisions are shown in **Figure 1**.

4.2 State transitions and rewards

Transitions among states in the created MDP model of the information system rely on decisions and there are two main probabilities P_1 and P_2. P_1 is the probability of the transition from state 1 (no attacker's connection) to state 2 (connected). P_2 is the probability of the transition from state 2 to state 3 (detected). There are no transitions from state 1 to state 3 directly and no transitions from the state 3 to the state 2. The probability of a transition from state



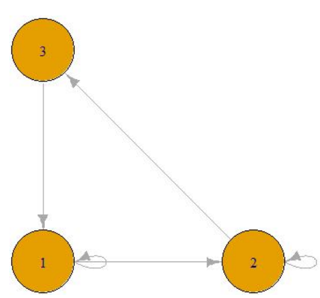


Figure 1. State transitions of two decisions: (a) decision 1 (wait) and (b) decision 2 (expel).

3 to state 1 is 0 for decision 1 and 1 for decision 2. The probability matrix of state transitions P_d and the reward matrix R_d for the two decisions are expressed as follows:

1) P_d and R_d for decision 1 are:

$$P_d = \begin{bmatrix} 1-P_1 & P_1 & 0 \\ 0 & 1-P_2 & P_2 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

$$R_d = \begin{bmatrix} 0 & r_{12} & 0 \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} = \begin{bmatrix} 0 & -C_a & 0 \\ 0 & -C_a & B_i - C_a \\ 0 & 0 & B_i - C_a \end{bmatrix} \tag{4}$$

where C_a is the cost due to attacking and B_i is the defender's benefit due to collecting information during the learning process of knowing about the attack.

2) P_d and R_d for decision 2 are:

$$P_d = \begin{bmatrix} 1-P_1 & P_1 & 0 \\ 0 & 1-P_2 & P_2 \\ 1 & 0 & 0 \end{bmatrix} \tag{5}$$

$$R_d = \begin{bmatrix} 0 & r_{12} & 0 \\ 0 & r_{22} & r_{23} \\ r_{31} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -C_a & 0 \\ 0 & -C_a & B_i - C_a \\ -C_e & 0 & 0 \end{bmatrix} \tag{6}$$

where C_e is the cost due to expelling.

5. Data analytics of the information system based on the MDP

5.1 Analytics based on MDP over an infinite planning horizon

Let $P_1 = 0.15$, $P_2 = 0.15$, $C_e = 1$, $B_i = 3$, $C_a = 5$. The analytics of the information system with a discount $\gamma = 0.85$ over an infinite planning horizon is conducted. Policy iteration and value iteration are used in the data analytics and the obtained optimal policies in both the two methods are d (1, 1, 2), indicating that decision 1, decision 1, and decision 2 are made on the state 1, the state 2, and the state 3, respectively. The total expected costs of the two methods and Q-learning are listed in **Table 1** to evaluate the model validity in this paper. Gauss-Seidel's algorithm is employed in VI for an improved convergence speed. The accuracy is also improved compared with the result of Jacob's algorithm. In Q-learning, the learning rate β is set to $1/\sqrt{n+2}$ in this paper and N is the number of iterations to perform. The results of policy iteration and the Gauss-Seidel method are the same and are close to that of Q-learning, which indicates the parameters in the MDP model are reasonable, and the created model is valid.

Table 1. Total expected costs of three states in the information system based on various algorithms over an infinite planning horizon ($\gamma = 0.85$).

Algorithms	C_1	C_2	C_3
VI (Jacob' algorithm)	12.68322	21.77953	11.77344
VI (Gauss-Seidel's algorithm)	12.73186	21.82816	11.82208
PI	12.73186	21.82816	11.82208
Q-learning (N = 120,000)	12.67515	21.63394	11.90482

5.2 Analytics over a finite planning horizon

The total expected costs of three states (states 1-3) are calculated utilizing the VI algorithm over a 40-step planning horizon with and without a discount, respectively. The rewards (the negative values of the costs in this paper) at the end of the planning horizon are set to 0 for three states for the beginning of the backward recursion of the VI. **Table 2** and **Table 3** show the computation results. $C_1(n)$, $C_2(n)$, and $C_3(n)$ represent the total expected cost at step n for the state 1, the state 2, and the state 3, respectively. It is shown that the total expected costs $C_1(n)$, $C_2(n)$, and $C_3(n)$ in **Table 2** are very close to C_1, C_2, and C_3 for infinite planning horizon in **Table 1**, respectively when Epoch $n \leq 10$ for a 40-step planning horizon ($\gamma = 0.85$).

5.3 Analytics of the information system with various parameters of the transition probability

Analytics of the information system with various state transition probability parameters P_1 and P_2 is performed based on the PI over an infinite planning horizon. The following data are utilized: $P_2 = 0.15$, $C_e = 1$, $B_i = 3$, $C_a = 5$, and $\gamma = 0.85$. The total expected cost C_i (i = 1, 2, 3) for states 1-3 at various P_1 is analyzed and the result is shown in **Figure 2**. All the values of C_1, C_2, and C_3 are increased with the increase of P_1.

Table 2. Total expected costs of three states computed using the VI algorithm over a 40-step planning horizon ($\gamma = 0.85$).

Epoch *n*	$C_1(n)$	$C_2(n)$	$C_3(n)$
0	12.7065	21.8028	11.7967
5	12.6746	21.7710	11.7649
10	12.6029	21.6992	11.6931
15	12.4412	21.5375	11.5315
20	12.0769	21.1731	11.1672
25	11.2565	20.3509	10.3471
30	9.4191	18.4836	8.5165
33	7.3930	16.3143	6.5226
35	5.4787	14.0296	4.6918
36	4.3432	12.4763	3.6503
37	3.1180	10.5134	2.5912
38	1.8720	7.9649	1.6375
39	0.75	4.55	1.00
40	0	0	0

Table 3. Total expected costs of three states computed using the VI algorithm over a 40-step planning horizon ($\gamma = 1.0$).

Epoch *n*	$C_1(n)$	$C_2(n)$	$C_3(n)$
0	85.3155	93.7710	76.7897
5	75.1760	83.7475	66.7897
10	64.9085	73.6947	56.7897
15	54.4104	63.5756	46.7897
20	43.5248	53.3072	36.7897
25	32.0626	42.7023	26.7897
30	19.9701	31.3391	16.7897
33	12.6354	23.7993	10.7897
35	7.9649	18.2667	6.7897
36	5.7897	15.2911	4.7946
37	3.7946	12.0945	3.0700
38	2.0700	8.5675	1.7500
39	0.75	4.55	1.00
40	0	0	0

Let $P_1 = 0.15$, $C_e = 1$, $B_i = 3$, $C_a = 5$, and $\gamma = 0.85$. The PI over an infinite planning horizon is utilized. The total expected cost C_i ($i = 1, 2, 3$) at various P_2 is shown in **Figure 3**. All the values of C_1, C_2, and C_3 are decreased with the increase of P_2.

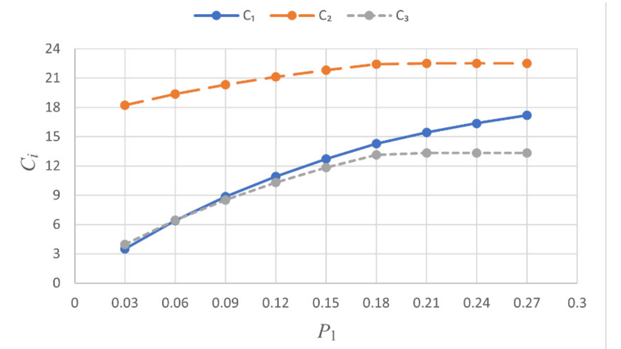


Figure 2. Total expected cost C_i ($i = 1, 2, 3$) at various P_1.

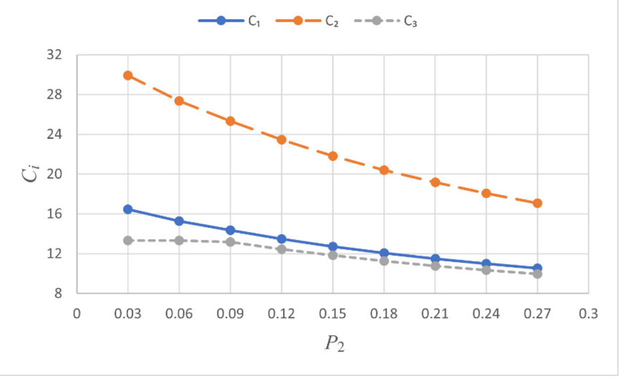


Figure 3. Total expected cost C_i ($i = 1, 2, 3$) at various P_2.

5.4 Analytics of the information system with various transition cost parameter C_a

Analytics of the information system with various transition cost parameters C_a is performed based on the PI over an infinite planning horizon. The following data are used: $P_1 = 0.15$, $P_2 = 0.15$, $C_e = 1$, $B_i = 3$, and $\gamma = 0.85$. **Figure 4** illustrates the total expected cost C_i ($i = 1, 2, 3$) at various C_a. The greater the value of C_a, the larger the value of the expected total cost C_i.

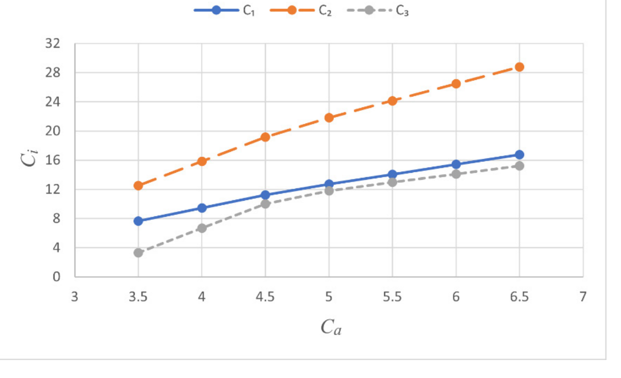


Figure 4. Total expected cost C_i ($i = 1, 2, 3$) at various C_a.

6. Data analytics of the information system based on POMDP

6.1 Analytics based on the POMDP over an infinite planning horizon

Analytics of the information system is performed based on a discounted POMDP over an infinite planning horizon. The following data are utilized: $P_1 = 0.15$, $P_2 = 0.15$, $C_e = 1$, $B_i = 3$, $C_a = 5$, and $\gamma = 0.85$. The following solution methods or algorithms are used in solving the POMDP problem: "grid", "enum", "twopass", "witness", "incprune", and "SARSOP" [23]. The total expected cost C_t is shown in **Table 4**, indicating that the result of SARSOP is very close to the results of the other five methods (with the same results).

6.2 The effects of various parameters on POMDP solutions

The following data are used to study the effects of various parameters on the total expected cost C_t: $C_e = 1$, $B_i = 3$, and $\gamma = 0.85$. **Figure 5** shows the effect of the connecting probability P_1 on C_t at various P_2 (0.03, 0.15, and 0.27) when $C_a = 5$. **Figure 6** shows the effect of P_1 on C_t at various C_a (3.5, 5.0, and 6.5) when $P_2 = 0.15$. It is shown that C_t is increased with an increase of P_1. Similarly, the effects of the detecting probability P_2 on the total expected cost C_t are studied. The results are shown in **Figure 7 and Figure 8**. It is shown that C_t is decreased with an increase of P_2. **Figure 9** shows the effect of C_a on C_t at various P_1 (0.03, 0.15, and 0.27) when $P_2 = 0.15$. **Figure 10** shows the effect of C_a on C_t at various P_2 (0.03, 0.15, and 0.27) when $P_1 = 0.15$. It is shown that C_t is increased with the increase of C_a.

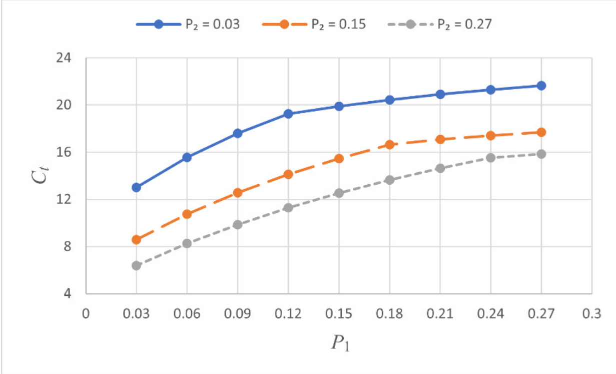


Figure 5. The effect of P_1 on C_t at various P_2 when $C_a = 5$.

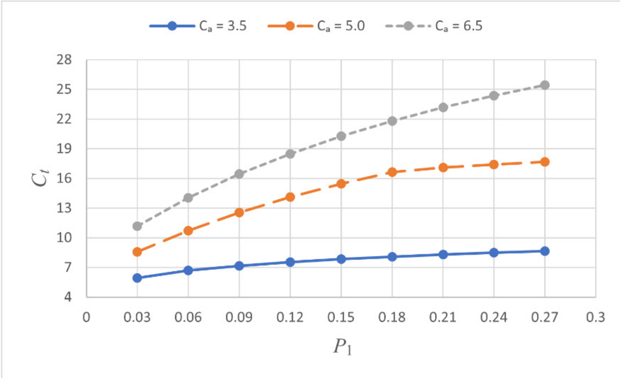


Figure 6. The effect of P_1 on C_t at various C_a when $P_2 = 0.15$.

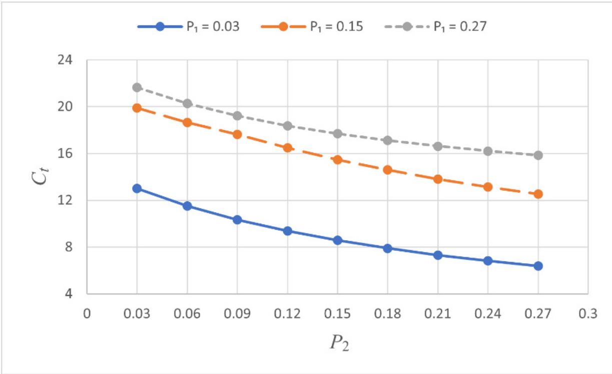


Figure 7. The effect of P_2 on C_t at various P_1 when $C_a = 5.0$.

Table 4. The total expected cost C_t based on six various methods.

Methods	**grid**	**enum**	**twopass**	**witness**	**incprune**	**SARSOP**
C_t	15.46070	15.46070	15.46070	15.46070	15.45570	15.46073

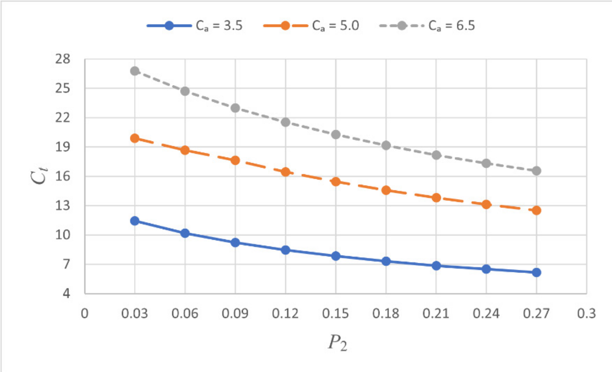


Figure 8. The effect of P_2 on C_t at various C_a when $P_1 = 0.15$.

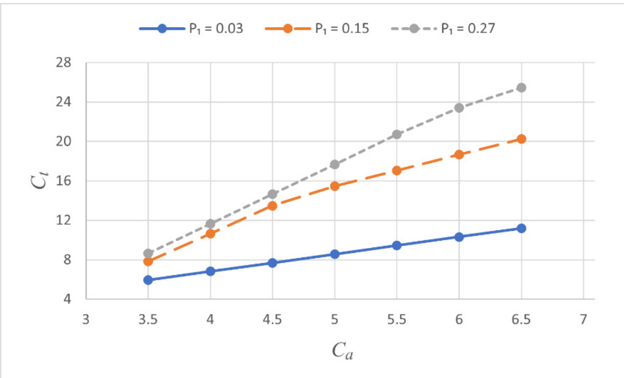


Figure 9. The effect of C_a on C_t at various P_1 when $P_2 = 0.15$.

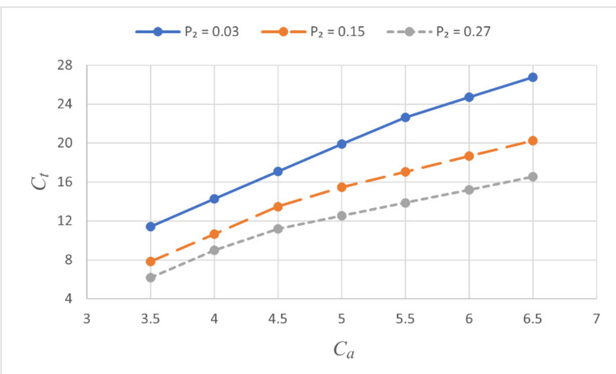


Figure 10. The effect of C_a on C_t at various P_2 when $P_1 = 0.15$.

7. Conclusions

Data analytics of an information system based on the MDP demonstrates that the algorithms in this paper are effective in achieving optimal policies to minimize the total expected costs of states of the information system. These algorithms are effective in analytics over a finite planning horizon and an infinite planning horizon (for a discounted MDP). The VI (Gauss-Seidel's algorithm) and the PI achieve the same results, and the result of Q-learning is very close to the results of the VI and the PI, indicating the MDP model is valid. The pros of data analytics of the information system based on the MDP lie in: 1) Multiple methods can be used to check the validity of the created MDP model; 2) It is convenient to perform predictive modelling and study the effects of various parameters on the total expected cost of the information system.

One of the main cons of the MDP-based method is that the state uncertainty is not considered while this problem is fixed in the POMDP method. In the analytics of a discounted POMDP (over an infinite planning horizon) of the information system, the total expected cost of the information system is increased with an increase in the connecting probability and is decreased with an increase in the detecting probability. The cost caused by the attacker is a primary factor in increasing the total expected cost of the information system.

Conflict of Interest

There is no conflict of interest.

Funding

This research received no external funding.

Acknowledgement

This paper is based upon work supported by Mississippi State University, USA.

References

[1] AlSadhan, T., Park, J.S. (editors), 2021. Leveraging information security continuous monitoring to enhance cybersecurity. 2021 International Conference on Computational Science and Computational Intelligence (CSCI); 2021 Dec 15-17; Las Vegas, NV, USA. USA: IEEE. p. 753-759.

[2] United States Cyber Command, 2018. Achieve and Maintain Cyberspace Superiority, Command Vision for U.S. Cyber Command [Internet]. Available from: https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20 Vision%20April%20 2018.pdf?ver=2018-06-14-152556-010

[3] Wendt, D., 2019. Addressing both sides of the cybersecurity equation. Journal of the Cyber Security & Information Systems Information Analysis Center. 7(2).

[4] Anuar, N.B., Sallehudin, H., Gani, A., et al., 2008. Identifying false alarm for network intrusion detection system using hybrid data mining and decision tree. Malaysian Journal of Computer Science. 21(2), 101-115.

[5] Kukielka, P., Kotulski, Z. (editors), 2008. Analysis of different architectures of neural networks for application in intrusion detection systems. 2008 International Multiconference on Computer Science and Information Technology; 2008 Oct 20-22; Wisla, Poland. USA: IEEE. p. 807-811.

[6] Faisal, M.A., Aung, Z., Williams, J.R., et al., 2012. Securing advanced metering infrastructure using intrusion detection system with data stream mining. In: Chau, M., Wang, G.A., Yue, W.T., et al. (editors), intelligence and security informatics. PAISI 2012. Lecture Notes in Computer Science. Springer, Berlin: Heidelberg. pp. 96-111.
DOI: https://doi.org/10.1007/978-3-642-30428-6_8

[7] Raiyn, J., 2014. A survey of cyber attack detection strategies. International Journal of Security and Its Applications. 8(1), 247-256.

[8] Singh, J., Nene, M.J., 2013. A survey on machine learning techniques for intrusion detection systems. International Journal of Advanced Research in Computer and Communication Engineering. 2(11), 4349-4355.

[9] Cardenas, A.A., Manadhata, P.K., Rajan, S.P., 2013. Big data analytics for security. IEEE Security & Privacy. 11(6), 74-76.

[10] Wiafe, I., Koranteng, F.N., Obeng, E.N., et al., 2020. Artificial intelligence for cybersecurity: A systematic mapping of literature. IEEE Access. 8, 146598-146612.

[11] Bao, N., Musacchio, J., 2009. Optimizing the decision to expel attackers from an information system. 2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton); 2009 Sep 30-Oct 2; Monticello, IL, USA. USA: IEEE. p. 644-651.

[12] Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. Foundations of machine learning. Adaptive computation and machine learning. MIT Press: USA.

[13] Alsheikh, M.A., Hoang, D.T., Niyato, D., et al., 2015. Markov decision processes with applications in wireless sensor networks: A survey. IEEE Communications Surveys & Tutorials. 17(3), 1239-1267.

[14] Chen, Y., Hong, J., Liu, C.C., 2018. Modeling of intrusion and defense for assessment of cyber security at power substations. IEEE Transactions on Smart Grid. 9(4), 2541-2552.

[15] van Otterlo, M., Wiering, M., 2012. Reinforcement learning and Markov decision processes. Reinforcement Learning. Springer, Berlin: Heidelberg. pp. 3-42.

[16] Sutton R.S., Barto, A.G., 2018. Reinforcement learning: An introduction. MIT press: USA.

[17] Zanini, E., 2014. Markov Decision Processes [Internet]. Available from: https://web.archive.org/web/20170812131743id_/http://www.lancs.ac.uk/~zaninie/MDP.pdf

[18] Liu, D., Khoukhi, L., Hafid, A. (editors), 2017. Data offloading in mobile cloud computing: A Markov decision process approach. 2017 IEEE International Conference on Communications (ICC); 2017 May 21-25; Paris, France. USA: IEEE. p. 1-6.

[19] Xiang, X., Foo, S., 2021. Recent advances in deep reinforcement learning applications for solving partially observable Markov decision processes (POMDP) problems: Part 1—fundamentals and applications in games, robotics and natural language processing. Machine Learning and Knowledge Extraction. 3(3), 554-581.

[20] Russell, S.J., Norvig, P., 2021. Artificial intelligence a modern approach, 4th edition. Pearson Education, Inc: UK.

[21] Kurniawati, H., Hsu, D., Lee, W.S., 2008. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. Robotics: Science and systems. MIT Press: USA. pp. 65-72.

[22] Cassandra, A.R., Kaelbling, L.P., Littman, M.L., 1994. Acting optimally in partially observable stochastic domains. Aaai. 94, 1023-1028.

[23] Kamalzadeh, H., Hahsler, M., 2019. POMDP: Introduction to Partially Observable Markov Decision Processes.