American Journal of Engineering and Applied Sciences

Original Research Paper

Data Analytics for the Cyber Security of an Information System Based on a Markov Decision Process Model

Lidong Wang, Randy Jones and Terril C. Falls

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

Article history
Received: 21-09-2022
Revised: 14-10-2022
Accepted: 18-10-2022

Corresponding Author:
Lidong Wang
Institute for Systems Engineering Research,
Mississippi State University,
United States
Email: lidong@iser.msstate.edu

Abstract: Intrusion detection is an important research topic in information systems and cyber security. Both a defender and an attacker detect and learn about each other during an intrusion process. The defender can expel the attacker as soon as the attacker is detected or wait and observe to know more about the attacker for the detection and prevention of other attacks in the future. An optimal decision is often required in this situation. Data analytics is conducted to achieve an optimal decision for the cyber security of an information system based on a Markov Decision Process (MDP) model in this study. The state of the information system is completely observable in the model. The model is validated using various algorithms that include policy iteration, value iteration, and Q-learning. Data analytics over a finite planning horizon and an infinite planning horizon is conducted, respectively. The expected total cost for each state is analyzed at various parameters of the transition probability and various parameters of the transition cost.

Keywords: Cyber Security, Information System, Markov Decision Process, Data Analytics, Q-Learning

Introduction

Intrusion detection and response is a basic component of network security. Intrusion Detection Systems (IDS) are significant elements for critical infrastructure security (Kiennert *et al*., 2019). Detection and prevention of attacks are generally more important than actions after attacks (Srujana *et al*., 2022). It is often a challenge to extract quality information for identifying exploited, infected, or vulnerable assets and taking suitable actions because cyber security observations over a network need to be extracted from big data that are frequently uncertain, noisy, and incomplete. A collaborative approach has been developed that integrates logistic regression, a partially observable Markov decision process, and online data analytics on temporal causality and dependency relationships of observations for identifying and controlling infection (Cam, 2017). The development of a security model for the dynamic defense of networks has been presented; it modeled interactions between exploits and security conditions (Miehling *et al*., 2017).

There are also challenges in monitoring insiders' behaviors. For example, collecting and analyzing massive logs are challenges for log auditing systems. The hidden Markov chain was introduced in a log auditing system for recording behaviors of users in the chain of time series; an improved hidden Markov model was proposed to construct a dynamic transformation of network behaviors. The accuracy of the algorithm based on the improved model has been increased and the performance of the overall audit system has been improved (Liu *et al*., 2018).

The combination of intrusion detection and continuous user authentication is an effective method of improving security performance in high-security mobile ad hoc networks (MANETs). A distributed optimal scheme for the combination of intrusion detection and user authentication has been developed (Bu *et al*., 2011). Intrusion detection was modeled as sensors to identify the system security state while multi-modal biometrics were utilized for authentication. The whole system was formulated as a Partially Observed Markov Decision Process (POMDP); hidden Markov model scheduling algorithms that are based on the dynamic programming method were used to derive an optimal scheme (Liu *et al*., 2009). An issue of mobile data offloading with an architecture of mobile cloud computing was studied. Mobile data were delivered by Wi-Fi or cellular and Device-to-Device (D2D) communication networks in the study. Part of cellular data traffic was offloaded through the D2D network and WiFi. The issue of data offloading was formulated as an MDP with a finite horizon. The issue was solved at a minimal total cost using a hybrid offloading algorithm (Liu *et al*., 2017).

A game-theoretic method deals with both a defender's and an attacker's behaviors in the analytics of security. It also provides a high-level direction for the overall optimization of IDS and the optimization lies in three aspects: IDS configuration, defense resource allocation, and countermeasure selection (Kiennert *et al*., 2019). An approach to automatic intrusion response, called the response and recovery engine, has been proposed based on the strategy of a game-theoretic response. Optimal actions of network-level response were decided through a game-theoretic optimization process in which the RRE tried to maximize its benefits (Zonouz *et al*., 2013). Both the defender and the attacker are learning about each other. The defender can expel the attacker as soon as the attacker is detected or wait to know more about the attacker. The more knowledge the defender achieves, the easier the defender to prevent the attacker now or in the future. The knowledge of a defender can be the attacker's objectives, attack methods, estimated technical level of the attacker, etc., (Bao and Musacchio, 2009). An optimal decision or action is often required in the intrusion process that indicates whether to expel the attacker and when to expel it.

The objective of this research is to conduct data analytics and achieve optimal decisions for the cyber security of an information system based on an MDP model, validate the model by comparing analytics results obtained using various algorithms and predict analytics results using various parameters of the transition probability and the transition reward or the transition cost. The *R* language and its functions are used to help data analytics.

Materials and Methods

Markov Decision Process and Algorithms

An MDP is well-defined using a tuple <*S*, *A*, *P*, *R*, γ> (Mohri *et al*., 2012; Alsheikh *et al*., 2015; Chen *et al*., 2016): *S* is the set of states; *A* refers to the set of actions; *P* represents the matrix of the transition probability that expresses a transition from the state *s* to the state *s'*($s \in S, s' \in S$) after the action *a* ($a \in A$); *R* refers to the immediate reward due to the action *a*; and γ ($0 < \gamma < 1$) is the discounted factor of the reward. Solving an MDP is often the process of finding optimal actions or an optimal policy to minimize the expected total cost or maximize the expected total reward.

Policy Iteration (PI), Value Iteration (VI), and Q-learning are often utilized to find an optimal policy for an MDP. Analytics results based on the algorithms of the three methods are often remarkably different or there is a convergence problem during iterations if the created MDP model is not practical due to unsuitable model parameters or an incorrect model structure. Thus, the three methods are employed in this research; results are compared to verify whether the MDP model is valid or not.

PI tries to find a better policy compared to the previous one. The iterative process of policy evaluation and policy improvement is stopped until two continuous policy iterations result in the same policy, indicating that an optimal policy is achieved. The policy iteration is described in Algorithm 1 (Otterlo and Wiering, 2012; Sutton and Barto, 2018). $P(s, a, s')$ is the transition probability. $R(s, a, s')$ is the immediate reward due to a transition from state *s* to state *s'* after action *a*. $V(s)$ and $V(s')$ are the expected total reward of *s* and *s'*, respectively. $\pi(s)$ is an optimal policy of *s*. $V(s)$ is calculated using the equation: $V(s) = \max_a \sum_{s'} P(s,\pi(s),s')\big(R(s,\pi(s),s') + \gamma V(s')\big)$. Δ is the value difference between two successive iterative steps. τ is the tolerance (a very small positive number).

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 PI (policy evaluation)
Repeat
$\Delta \leftarrow 0$
For each $s \in S$
$v \leftarrow V(s)$
$V(s) \leftarrow \max_a \sum_{s'} P(s,\pi(s),s')(R(s,\pi(s),s') + \gamma V(s'))$
$\Delta \leftarrow \max\left(\Delta, |V(s) - v|\right)$
Until $\Delta < \tau$
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 a policy is stable, then stop; else go to step 2

The optimal policy of the MDP can also be achieved based on VI (Otterlo and Wiering, 2012; Zanini, 2014). VI of each state uses the following equation to compute $V(s)$: $V(s) = \max_a \sum_{s'} P(s,\pi(s),s')\big(R(s,\pi(s),s') + \gamma V(s')\big)$. A stopping criterion for VI is employed to assess the convergence during iterations. The criterion is: The value difference Δ between two successive steps of iterations is less than the tolerance τ. Algorithm 2 shows the value iteration process.

Q-learning (Liu *et al*., 2017; Zanini, 2014) allows an agent to learn a Q-value function that 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 study. It can be used to perform data analytics and simulation of an MDP with the discounted value of γ 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 the episode. The best action at state s is chosen according to the optimal policy $\pi(s)$.

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 $s \in S$
$v \leftarrow V(s)$
$V(s) \leftarrow \max_a \sum_{s'} P(s, \pi(s), s')(R(s, \pi(s), s') + \gamma V(s'))$
$\Delta \leftarrow \max(\Delta, |V(s)-v|)$
until $\Delta < \tau$
3 Output the optimal policy & the maximal $V(s)$.

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 the episode.
3 Output the optimal policy & the maximal $V(s)$.

An MDP Model of an Information System and the Structure of the Model

The information system has the following states: State 1—no attacker is connected to the system; state 2—an attacker is connected to the 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. After an expelling action, the system will return to state 1.

We have created an MDP model of the information system. The state transitions of the three states of the two decisions are shown in Fig. 1.

State Transitions and Rewards

A transition between states in the MDP of the information system depends upon the decision and there are two main probabilities p_{12} and p_{23}. p_{12} is the probability of a transition from state 1 (no attacker's connection) to state 2 (connected). p_{23} is the probability of a transition from state 2 to state 3 (detected). There is not any transition from state 1 to state 3 directly; there is not any transition from state 3 to state 2. The transition probability from state 3 to state 1 is 0 for decision 1; it is 1 for decision 2. The probability matrix of the state transition 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:

$$P_d = \begin{bmatrix} 1-p_{12} & p_{12} & 0 \\ 0 & 1-p_{23} & p_{23} \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

$$R_d = \begin{bmatrix} 0 & r_{12} & 0 \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} = \begin{bmatrix} 0 & c_{12} & 0 \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{bmatrix} \quad (2)$$

where, r_{12}, r_{22}, r_{23}, and r_{33} represent a reward (a negative value of a cost in this study) due to a transition from one state to another one, respectively. c_{12} is the cost due to the transition from state 1 to state 2 and c_{23} is the cost from state 2 to state 3; c_{22} and c_{33} are costs due to self-transitions. 0 indicates zero cost for a self-transition or no-state transition (Fig. 1).

2) P_d and R_d for decision 2 are:

$$P_d = \begin{bmatrix} 1-p_{12} & p_{12} & 0 \\ 0 & 1-p_{23} & p_{23} \\ 1 & 0 & 0 \end{bmatrix} \quad (3)$$

$$R_d = \begin{bmatrix} 0 & r_{12} & 0 \\ 0 & r_{22} & r_{23} \\ r_{31} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -c_{12} & 0 \\ 0 & -c_{22} & -c_{23} \\ -c_{31} & 0 & 0 \end{bmatrix} \quad (4)$$

where, c_{31} is the cost due to a transition from state 3 to state 1.

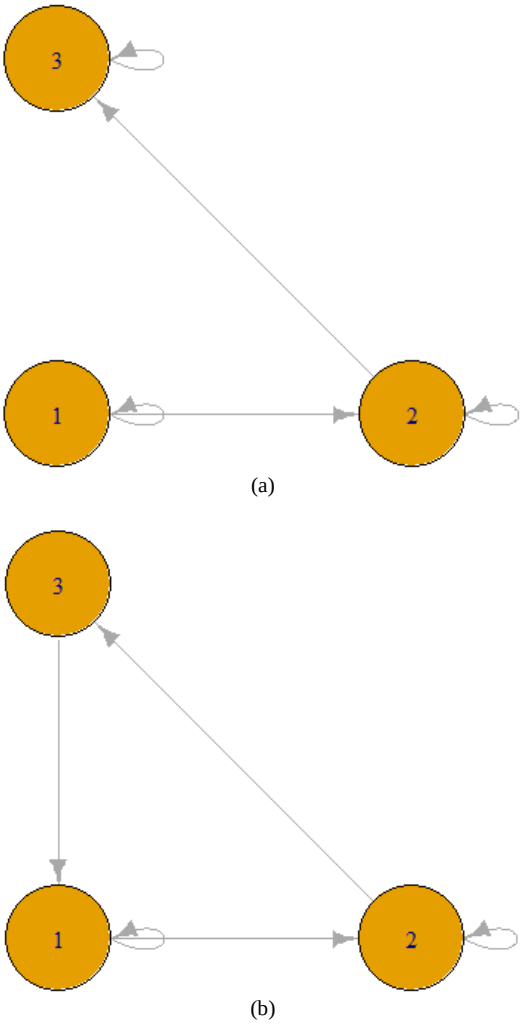

Fig. 1: State transitions of two decisions: (a) decision 1 (wait) and (b) decision 2 (expel)

Results

Data Analytics over a Finite Planning Horizon

Let $p_{12} = 0.2$, $p_{23} = 0.3$, $c_{12} = 1$, $c_{22} = 3$, $c_{23} = 2.5$, $c_{33} = 4$, $c_{31} = 2$. Substitute the data into Eq. (1)-(4), and the values of P_d and R_d for various decisions can be computed.

P_d and R_d for decision 1 are:

$$P_d = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0 & 0.7 & 0.3 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R_d = \begin{bmatrix} 0 & -1 & 0 \\ 0 & -3 & -2.5 \\ 0 & 0 & -4 \end{bmatrix}$$

P_d and R_d for decision 2 are:

$$P_d = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0 & 0.7 & 0.3 \\ 1 & 0 & 0 \end{bmatrix}$$

$$R_d = \begin{bmatrix} 0 & -1 & 0 \\ 0 & -3 & -2.5 \\ -2 & 0 & 0 \end{bmatrix}$$

Expected total costs of the states are calculated using the VI algorithm over a 6-step horizon with and without a discount, respectively. The rewards (the negative values of the costs in this study) at the end of the horizon are set to zero for states for the beginning of the backward recursion of the value iteration. Computation results are shown in Tables 1 and 2. $C1(n)$, $C2(n)$, and $C3(n)$ are the expected total cost of state 1, state 2, and state 3 at step n, respectively. The calculated optimal policy is d (1, 1, 2), indicating that decision 1, decision 1, and decision 2 are made on state 1, state 2, and state 3, respectively. The data analytics in this study is finished using *R* language.

Data Analytics over an Infinite Planning Horizon

The above data ($p_{12} = 0.2$, $p_{23} = 0.3$, $c_{12} = 1$, $c_{22} = 3$, $c_{23} = 2.5$, $c_{33} = 4$, $c_{31} = 2$) are utilized too in the analytics of the information system with $\gamma = 0.9$ over an infinite planning horizon. PI and VI are used in the data analytics and the obtained optimal policies in both two methods are d (1, 1, 2). A comparison of the expected total costs of the two methods and Q-learning is shown in Table 3 to verify whether the MDP model in this study is valid or not.

Table 1: Expected total costs of three states obtained using a VI algorithm over a 6-step planning horizon (no discount, $\gamma = 1$)

Epoch n	$C1(n)$	$C2(n)$	$C3(n)$
0	5.934	11.618	6.603
1	4.603	10.258	5.291
2	3.291	8.854	4.033
3	2.033	7.322	2.930
4	0.930	5.445	2.200
5	0.200	2.850	2.000
6	0.000	0.000	0.000

Table 2: Expected total costs of three states calculated using the VI algorithm over a 6-step planning horizon (the discount $\gamma = 0.9$)

Epoch n	$C1(n)$	$C2(n)$	$C3(n)$
0	4.314	9.547	5.176
1	3.529	8.744	4.401
2	2.667	7.823	3.575
3	1.750	6.705	2.771
4	0.857	5.186	2.180
5	0.200	2.850	2.000
6	0.000	0.000	0.000

Gauss-Seidel's algorithm is employed in VI for an improved convergence speed. In Q-learning, the learning rate β is set to $1/\sqrt{n+2}$ in this study; n is the number of iterations to perform. The results of PI and VI are almost the same and they are close to the results of Q-learning, which indicates that the parameters of the MDP are suitable and the MDP model is valid.

Analytics of the Information System with Various Parameters of the Transition Probability

Analytics over an infinite horizon with various parameters of the state transition probability p_{12} and p_{23} is performed based on policy iteration. The following data are utilized in the analytics: $c_{12} = 1$, $c_{22} = 3$, $c_{23} = 2.5$, $c_{33} = 4$, $c_{31} = 2$, $\gamma = 0.9$ and $p_{23} = 0.3$. Expected total cost $C = (C1, C2, C3)$ for the three states at various p_{12} (0.1, 0.2, 0.3, 0.4, and 0.5) is analyzed and the result is shown in Fig. 2. All the values of $C1$, $C2$ and $C3$ are increased and $C1$ and $C3$ become very close with the increase of p_{12}.

Let $c_{12} = 1$, $c_{22} = 3$, $c_{23} = 2.5$, $c_{33} = 4$, $c_{31} = 2$, $\gamma = 0.9$ and $p_{12} = 0.2$. Expected total cost $C = (C1, C2, C3)$ at various p_{23} (0.1, 0.2, 0.3, 0.4, and 0.5) is shown in Fig. 3. All the values of $C1$, $C2$, and $C3$ are increased and the difference between $C1$ and $C3$ becomes larger with the increase of p_{23}.

Analytics of the Information System with Various Parameters of the Transition Cost

Let $\rho = c_{22}/c_{12}$, $\mu = c_{23}/c_{12}$ and $\varphi = c_{31}/c_{12}$. Analytics over an infinite planning horizon with various parameters of the transition cost is performed based on policy iteration. The following data are used: $p_{12} = 0.2$, $p_{23} = 0.3$, $c_{12} = 1$, $c_{23} = 2.5$, $c_{33} = 4$, $c_{31} = 2$ and $\gamma = 0.9$. Expected total cost $C = (C1, C2, C3)$ at various ρ (1.0, 1.5, 2.0, 2.5, and 3.0) is shown in Fig. 4. The greater the value of ρ, the larger the value of expected total cost C.

Let $p_{12} = 0.2$, $p_{23} = 0.3$, $c_{12} = 1$, $c_{22} = 3$, $c_{33} = 4$, $c_{31} = 2$ and $\gamma = 0.9$. Expected total cost $C = (C1, C2, C3)$ at various s (2.0, 2.5, 3.0, 3.5, and 4.0) is shown in Fig. 5 which illustrates a similar trend to Fig. 4.

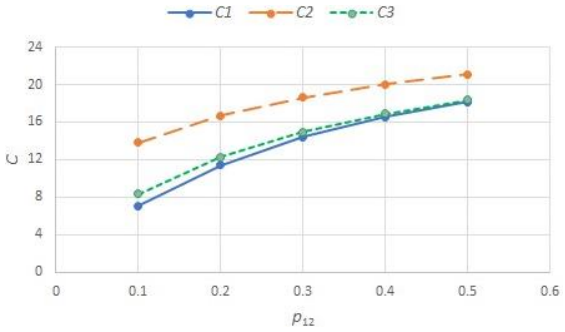

Fig. 2: Expected total cost C ($C1$, $C2$, $C3$) of three states at various p_{12}

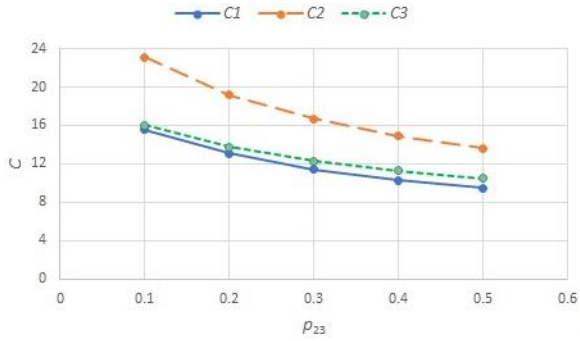

Fig. 3: Expected total cost C ($C1$, $C2$, $C3$) of three states at various p_{23}

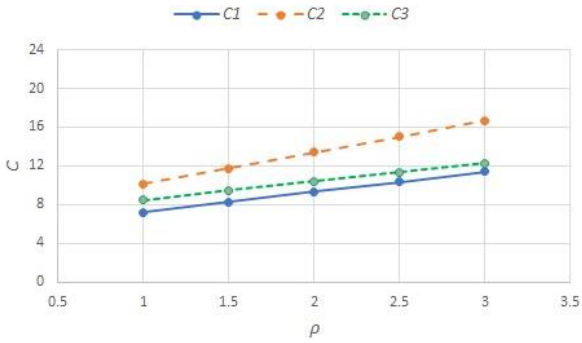

Fig. 4: Expected total cost C ($C1$, $C2$, $C3$) of three states at various ρ

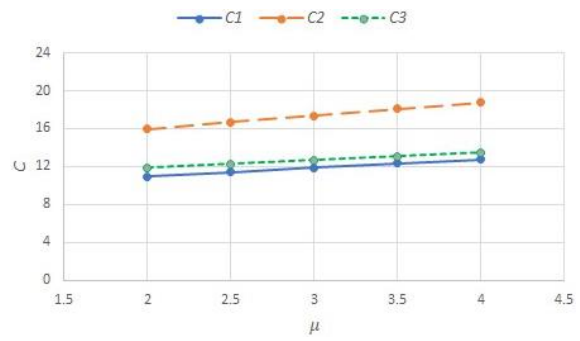

Fig. 5: Expected total cost C ($C1$, $C2$, $C3$) of three states at various μ

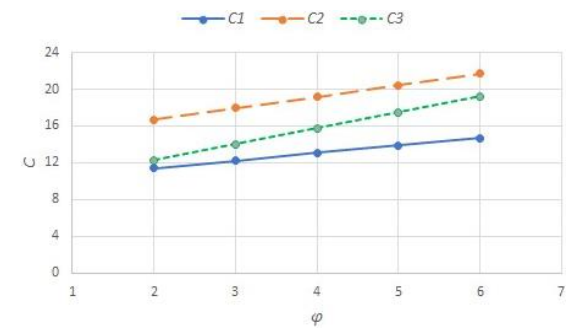

Fig. 6: Expected total cost C ($C1$, $C2$, $C3$) at various φ

Table 3: Expected total costs of three states over an infinite planning horizon in the information system based on various methods ($\gamma = 0.9$)

Methods	$C1$	$C2$	$C3$
PI	11.4300	16.6689	12.2870
VI (Gauss-Seidel's algorithm)	11.4299	16.6688	12.2869
Q-learning (n = 90000)	11.5869	16.5012	12.4182
Q-learning (n = 95000)	11.5025	16.5658	12.3215a

Let $p_{12} = 0.2$, $p_{23} = 0.3$, $c_{12} = 1$, $c_{22} = 3$, $c_{23} = 2.5$, $c_{33} = 4$, $\gamma = 0.9$. Expected total cost $C = (C1, C2, C3)$ at various φ (2, 3, 4, 5, and 6) is shown in Fig. 6. Each value of expected total cost C is increased with the increase of φ. $C3$ is close to $C1$ in the beginning, but it gradually becomes close to $C2$ when φ increases.

Discussion

Compared with Bao and Musacchio's published research and results, it is easier and more convenient to employ the methods in this study, perform data analytics and obtain desirable results using *R* language and its functions. It is easier to investigate the effects of various parameters on the expected total cost C ($C1$, $C2$, $C3$); therefore, the parameters are more controllable.

There are limitations to this study. Firstly, the state of the information system is completely observable in the created MDP model in this research. But in many real applications, some evidence may be deterministic while others may not be completely observable. This partially observable status leads to the uncertainty of the system state. For example, no evidence of an attack in the information system may mean two possible situations: (1) No attacker is connected to the system; (2) an attacker is not detected due to the limited capability of the defender and the system.

Secondly, five parameters are used to describe transition costs in the paper. They are: c_{12}, c_{22}, c_{23}, c_{31} and c_{33}. We will try to reduce the number of parameters, which will further simplify the MDP model. This is our ongoing research as well as future work.

Thirdly, a defender and an attacker try to learn more about each other during an intrusion process. The knowledge of the attacker regarding the defender and the information system is often increased with time during the intrusion process. The knowledge of the defender regarding the attacker helps to detect and expel the attacker in time, which will finally improve the cybersecurity of the information system. The evolution of their knowledge indicates the process of learning that is a dynamic process. The five parameters for transition costs and two main probabilities p_{12} and p_{23} are used in the paper to describe the learning process and the intrusion/prevention process. It is better to create an intelligent model with the capability of reinforcing knowledge representations and the function of quantifying the dynamic learning process.

Conclusion

Intrusion detection and intrusion prevention are very important components of the cybersecurity of an information system. The intrusion detection process is often a learning process that both a defender and an attacker detect and learn about each other. An optimal decision deals with whether to expel the attack and when to expel it to achieve a minimum of the expected total cost for each state of the information system. Data analytics for the cybersecurity of the information system has been completed based on the created MDP model using *R* language and its functions. The results of PI and VI (with Gauss-Seidel's algorithm) are almost the same and they are very close to the results of Q-learning (with various iteration step numbers). This demonstrates the validity of the model and the effectiveness of methods and algorithms in this research for achieving an optimal policy that minimizes the expected total cost. The algorithms are effective and efficient in the analytics over a finite planning horizon or an infinite horizon (for a discounted MDP) at various transition probabilities and transition costs.

Future topics include (1) analytics of the information system based on a partially observable Markov decision process (POMDP); (2) the reduction of the parameters of transition costs to further simplify the MDP model; and (3) further research on the knowledge evolution of the attack and the defender to create an intelligent model with the capability of reinforced knowledge representations and the function of quantifying the dynamic learning process.

Acknowledgment

The authors would like to thank Mississippi State University and the U.S. Army Engineer Research and Development Center for their support.

Disclaimer

Any opinions, findings, and conclusions, or recommendations in this research are those of the authors and do not reflect the views of the U.S. Army Engineer Research and Development Center.

Funding Information

This research is based on the work performed under Contract No. W912HZ-17-C-0015 of Mississippi State University with the U.S. Army ERDC.

Author's Contributions

All authors equally contributed to this study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all the other authors have read and approved the manuscript and no ethical issues are involved.

References

Alsheikh, M. A., Hoang, D. T., Niyato, D., Tan, H. P., & Lin, S. (2015). Markov decision processes with applications in wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials*, *17*(3), 1239-1267. https://doi.org/10.1109/COMST.2015.2420686

Bao, N., & Musacchio, J. (2009). Optimizing the decision to expel attackers from an information system. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* 2009 Sep 30 (pp. 644-651). IEEE. https://doi.org/10.1109/ALLERTON.2009.5394923

Bu, S., Yu, F. R., Liu, X. P., & Tang, H. (2011). Structural results for combined continuous user authentication and intrusion detection in high security mobile ad-hoc networks. *IEEE Transactions on Wireless Communications*, *10*(9), 3064-3073. https://doi.org/10.1109/TWC.2011.071411.102123

Chen, Y., Hong, J., & Liu, C. C. (2016). Modeling of intrusion and defense for assessment of cyber security at power substations. *IEEE Transactions on Smart Grid*, *9*(4), 2541-2552. https://doi.org/10.1109/TSG.2016.2614603

Kiennert, C., Ismail, Z., Debar, H., & Leneutre, J. (2019). A survey on game-theoretic approaches for intrusion detection and response optimization. *ACM Computing Surveys (CSUR)*, *51*(5), 1-31. https://doi.org/10.1145/3232848

Liu, C., Zhong, Y., & Wang, Y. (2018). Improved detection of user malicious behavior through log mining based on IHMM. In *2018 5th International Conference on Systems and Informatics (ICSAI)* 2018 Nov 10 (pp. 1193-1198). IEEE https://doi.org/10.1109/ICSAI.2018.8599422

Liu, D., Khoukhi, L., & Hafid, A. (2017). Data offloading in mobile cloud computing: A Markov Decision Process approach. In *2017 IEEE international conference on communications (ICC)* 2017 May 21 (pp. 1-6). IEEE. https://doi.org/10.1109/ICC.2017.7997070

Liu, J., Yu, F. R., Lung, C. H., & Tang, H. (2009). Optimal combined intrusion detection and biometric-based continuous authentication in high-security mobile ad hoc networks. *IEEE Transactions on Wireless Communications*, *8*(2), 806-815. https://doi.org/10.1109/TWC.2009.071036

Miehling, E., Rasouli, M., & Teneketzis, D. (2017). A dependency graph formalism for the dynamic defense of cyber networks. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* 2017 Nov 14 (pp. 511-512). IEEE. https://doi.org/10.1109/GlobalSIP.2017.8308695

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning: Adaptive computation and machine learning*. MIT Press. ISBN: 978-0262018258

Otterlo, M. V., & Wiering, M. (2012). *Reinforcement learning and Markov decision processes*. In *Reinforcement learning* (pp. 3-42). Springer, Berlin, Heidelberg. ISBN: 978-3-642-27645-3

Srujana, S., Sreeja, P., Swetha, G., & Shanmugasundaram, H. (2022). Cutting Edge Technologies for Improved Cybersecurity Model: A Survey. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 1392-1396), IEEE. https://doi.org/10.1109/ICAAIC53929.2022.9793228

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press. Cambridge, MA, 22447. ISBN: 978-0262039246

Zanini, E. (2014). Markov Decision Processes.

Zonouz, S. A., Khurana, H., Sanders, W. H., & Yardley, T. M. (2013). RRE: A game-theoretic intrusion response and recovery engine. *IEEE Transactions on Parallel and Distributed Systems*, *25*(2), 395-406. https://doi.org/10.1109/TPDS.2013.211