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 single
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 \(<S, A, P, R, \gamma>\). \(S\) 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 \(\gamma\) (\(0 < \gamma < 1\)) 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. \(P(s, a, s')\) 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. The stopping criterion is that the value difference of two successive iterative steps is less than the tolerance \(\tau\) (a very small positive number). Algorithm 2 shows the value iteration process.

Q-learning 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 \(\pi(s)\).
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:

\[
b'(s') = a P(o|s') \sum_s P(s'|s, a) b(s)
\]  

(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 \[22\]:

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

(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).](image-url)
3 to state 1 is 0 for decision 1 and 1 for decision 2. The probability matrix of state transitions \( P_a \) and the reward matrix \( R_d \) for the two decisions are expressed as follows:

1) \( P_a \) and \( R_d \) for decision 1 are:

\[
P_a = \begin{bmatrix}
1 - P_1 & P_1 & 0 \\
0 & 1 - P_2 & P_2 \\
0 & 0 & 1
\end{bmatrix}
\]

\[
R_d = \begin{bmatrix}
0 & r_{12} & 0 \\
r_{22} & 0 & r_{23} \\
0 & r_{32} & 0
\end{bmatrix} = \begin{bmatrix}
0 & -C_a & 0 \\
0 & -C_a & B_i - C_a \\
0 & 0 & B_i - C_a
\end{bmatrix}
\]

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) \( P_a \) and \( R_d \) for decision 2 are:

\[
P_a = \begin{bmatrix}
1 - P_1 & P_1 & 0 \\
0 & 1 - P_2 & P_2 \\
1 & 0 & 0
\end{bmatrix}
\]

\[
R_d = \begin{bmatrix}
0 & r_{12} & 0 \\
r_{22} & 0 & r_{23} \\
r_{31} & 0 & 0
\end{bmatrix} = \begin{bmatrix}
0 & -C_a & 0 \\
0 & -C_a & B_i - C_a \\
-C_e & 0 & 0
\end{bmatrix}
\]

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 \( \beta \) 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 (\( \gamma = 0.85 \)).

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI (Jacob’s algorithm)</td>
<td>12.68322</td>
<td>21.77953</td>
<td>11.77344</td>
</tr>
<tr>
<td>VI (Gauss-Seidel’s algorithm)</td>
<td>12.73186</td>
<td>21.82816</td>
<td>11.82208</td>
</tr>
<tr>
<td>PI</td>
<td>12.73186</td>
<td>21.82816</td>
<td>11.82208</td>
</tr>
<tr>
<td>Q-learning (( N = 120,000 ))</td>
<td>12.67515</td>
<td>21.63394</td>
<td>11.90482</td>
</tr>
</tbody>
</table>

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. \( C_i(n) \), \( C_2(n) \), 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 \( C_i(n) \), \( C_2(n) \), 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 ($\gamma = 0.85$).

<table>
<thead>
<tr>
<th>Epoch $n$</th>
<th>$C_1(n)$</th>
<th>$C_2(n)$</th>
<th>$C_3(n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12.7065</td>
<td>21.8028</td>
<td>11.7967</td>
</tr>
<tr>
<td>5</td>
<td>12.6746</td>
<td>21.7710</td>
<td>11.7649</td>
</tr>
<tr>
<td>10</td>
<td>12.6029</td>
<td>21.6992</td>
<td>11.6931</td>
</tr>
<tr>
<td>15</td>
<td>12.4412</td>
<td>21.5375</td>
<td>11.5315</td>
</tr>
<tr>
<td>20</td>
<td>12.0769</td>
<td>21.1731</td>
<td>11.1672</td>
</tr>
<tr>
<td>25</td>
<td>11.2565</td>
<td>20.3509</td>
<td>10.3471</td>
</tr>
<tr>
<td>30</td>
<td>9.4191</td>
<td>18.4836</td>
<td>8.5165</td>
</tr>
<tr>
<td>33</td>
<td>7.3930</td>
<td>16.3143</td>
<td>6.5226</td>
</tr>
<tr>
<td>35</td>
<td>4.3432</td>
<td>12.4763</td>
<td>3.6503</td>
</tr>
<tr>
<td>36</td>
<td>1.8720</td>
<td>7.9649</td>
<td>1.6375</td>
</tr>
<tr>
<td>37</td>
<td>0.75</td>
<td>4.55</td>
<td>1.00</td>
</tr>
<tr>
<td>38</td>
<td>0.75</td>
<td>4.55</td>
<td>1.00</td>
</tr>
<tr>
<td>39</td>
<td>0.75</td>
<td>4.55</td>
<td>1.00</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Epoch $n$</th>
<th>$C_1(n)$</th>
<th>$C_2(n)$</th>
<th>$C_3(n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>85.3155</td>
<td>93.7710</td>
<td>76.7897</td>
</tr>
<tr>
<td>5</td>
<td>75.1760</td>
<td>83.7475</td>
<td>66.5287</td>
</tr>
<tr>
<td>10</td>
<td>64.9085</td>
<td>73.6947</td>
<td>56.7897</td>
</tr>
<tr>
<td>15</td>
<td>54.4104</td>
<td>63.5756</td>
<td>46.7897</td>
</tr>
<tr>
<td>20</td>
<td>43.5248</td>
<td>53.3072</td>
<td>36.7897</td>
</tr>
<tr>
<td>25</td>
<td>32.0626</td>
<td>42.7023</td>
<td>26.7897</td>
</tr>
<tr>
<td>30</td>
<td>19.9701</td>
<td>31.3391</td>
<td>16.7897</td>
</tr>
<tr>
<td>33</td>
<td>12.6354</td>
<td>23.7993</td>
<td>10.7897</td>
</tr>
<tr>
<td>35</td>
<td>7.9649</td>
<td>18.2667</td>
<td>6.7897</td>
</tr>
<tr>
<td>36</td>
<td>5.7897</td>
<td>15.2911</td>
<td>4.7946</td>
</tr>
<tr>
<td>37</td>
<td>3.7946</td>
<td>12.0945</td>
<td>3.0700</td>
</tr>
<tr>
<td>38</td>
<td>2.0700</td>
<td>8.5675</td>
<td>1.7500</td>
</tr>
<tr>
<td>39</td>
<td>0.75</td>
<td>4.55</td>
<td>1.00</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 $P_2$.

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 $C_r$. 

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 $P_2$.

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”, “two-pass”, “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](image1)

Figure 5. The effect of \( P_1 \) on \( C_t \) at various \( P_2 \) when \( C_a = 5. \)

![Figure 6](image2)

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

![Figure 7](image3)

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

![Figure 8](image4)

Figure 8. 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>grid</th>
<th>enum</th>
<th>twopass</th>
<th>witness</th>
<th>incprune</th>
<th>SARSOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_t )</td>
<td>15.46070</td>
<td>15.46070</td>
<td>15.46070</td>
<td>15.46070</td>
<td>15.45570</td>
<td>15.46073</td>
</tr>
</tbody>
</table>
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


