Assessment of Reinforcement Learning-Based Penetration Testing Methodologies

2210117 CHEN Zhi

With the increasing complexity and frequency of cyber attacks, traditional penetration testing methods relying on manual operations are becoming increasingly inadequate. Penetration testing is a technique that simulates the behavior of an attacker to identify security vulnerabilities in a system. Traditional penetration testing often relies on manual operations, which are time-consuming and expensive, and the test results are highly dependent on the experience and skills of the testers.

To improve the efficiency and effectiveness of penetration testing, researchers are exploring the application of reinforcement learning in the field of penetration testing. Reinforcement Learning is a branch of machine learning, that involves an agent interacting with the environment and continuously learning the optimal strategy to maximize cumulative rewards. In penetration testing, reinforcement learning can automate the process of discovering and exploiting system vulnerabilities, thereby improving the coverage and depth of the tests. In recent years, many research efforts have demonstrated the potential of reinforcement learning in automated penetration testing.

This research report provides a comprehensive review and analysis of the application of reinforcement learning in automated penetration testing. We first conduct an in-depth analysis of around 30 relevant papers to identify the main problems and challenges in current research, as well as summarize the root causes and impacts of these issues. This process provides a reference for future research, helping to avoid repetitive mistakes and improve the effectiveness and precision of studies.

In addition, designing an evaluation method and criteria based on experimental results is essential for assessing the effectiveness of automated penetration testing. To this end, we develop a reliable and feasible evaluation method and standard based on our literature analysis. Furthermore, we conduct a series of experiments using the Network Attack Simulator (NAS) to ensure the reliability and credibility of the results.

The conclusions of this research demonstrate that while reinforcement learning has made significant progress in automated penetration testing, several critical challenges remain. These include the need for more scalable algorithms capable of handling large and complex network environments, the issue of sparse rewards which can hinder effective learning, and the need for greater adaptability to dynamic network changes.

Future work should focus on addressing these challenges through innovations in algorithm design, integration of domain knowledge, and development

of more realistic simulation environments. By overcoming these obstacles, reinforcement learning has the potential to revolutionize the field of penetration testing, making it more efficient, effective, and accessible to a broader range of cybersecurity professionals.
Keywords: Reinforcement Learning, Automated Penetration Testing, Literature Review