УДК 004.622, 004.056, 004.85

МЕТОДЫ ПРЕДОБРАБОТКИ ДАННЫХ ДЛЯ ИДЕНТИФИКАЦИИ УЯЗВИМОСТЕЙ В JAVA-ПРИЛОЖЕНИЯХ МЕТОДАМИ ГЛУБОКОГО ОБУЧЕНИЯ

Роенко Д.В. (Федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет ИТМО")

Научный руководитель – к.т.н., доцент (квалификационная категория "ординарный доцент") Менщиков А.А. (Федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет ИТМО")

В работе рассматриваются методы предобработки исходных кодов Java приложений для обнаружения уязвимостей методами глубокого обучения. Выбранные методы были использованы для подготовки собранного датасета кодовых фрагментов к дальнейшему анализу нейросетевыми алгоритмами.

Введение. Согласно последней статистике компании Positive Technologies за III квартал 2021 года, число инцидентов кибербезопасности по-прежнему остаётся на высоком уровне, среди которых около 11% напрямую связаны с эксплуатацией уязвимостей. Одним из недавних крупных инцидентов стала уязвимость в библиотеке Log4j, которой оказались подвержены миллионы Java-приложений. Она позволяла производить удалённое выполнение кода, что потенциально давало возможность захватить полный контроль над системой.

В настоящее время для поиска и последующего устранения ошибок безопасности используются сканеры уязвимостей, однако у них есть недостатки, такие как большое количество ложных срабатываний, ограниченность в определении ошибок безопасности из-за вручную задаваемых правил. Системы обнаружения уязвимостей на основе методов глубокого обучения должны быть лишены недостатков, присущих классическим методам анализа программ. Первоначальными этапами создания такой системы являются сбор датасета и его предобработка для последующего анализа алгоритмами глубокого обучения.

Основная часть. Для обучения классификатора можно использовать несколько источников данных. Juliet Test Suite и некоторые другие тестовые наборы SARD (Software Assurance Reference Dataset) относятся к синтетическим данным, они были искусственно созданы и размечены. Полусинтетические данные, содержащие реальные уязвимости, но изолированные от контекста возможно собрать из примеров кода базы данных CWE (Common Weakness Enumeration). Примеры кода, не являющиеся полностью или частично искусственными можно получить только из реальных проектов, например, с GitHub. Чем менее искусственным будет датасет, тем больше вероятность, что система идентифицирует уязвимость в реальном проекте.

На этапе предобработки данных первоначально предлагается компилировать исходный код участка программы в байткод Java, что позволяет обнаружить уязвимости в любых других языках, запускаемых на JVM (Java Virtual Machine). Далее байткод трансформируется в упрощённое представление Jimple, по которому строятся граф потока выполнения (CFG). Для каждой отдельной ветви этого графа выполняется токенизация и перевод в векторное представление

Выводы. В результате проделанной работы был произведён сбор данных, обзор методов их предобработки, с использованием выбранных методов, данные подготовлены для дальнейшего анализа алгоритмами глубокого обучения.

Роенко Д.В. (автор) Подпись

Менщиков А.А. (научный руководитель) Подпись