1. 004.02
2. Применение алгоритма градиентного бустинга в задаче идентификации программ по побайтовым сигнатурам
3. В.С. Багно
4. И.Е. Кривцова
5. Основные части тезиса

Краткое введение:

Объект исследования – идентификация программ по побайтовым сигнатурам. Предмет исследования – метод градиентного бустинга для решения задачи идентификации. Идентификация программ является одной из актуальных проблем информационной безопасности, поскольку возможность поиска вредоносных программ по их сигнатурам позволит быстро выявить и ликвидировать угрозу. В данной работе предложен способ идентификации программы по ее исходному коду, преобразованному в побайтовую сигнатуру, с помощью алгоритма градиентного бустинга.

Цели работы:

Изучить реализации алгоритма градиентного бустинга, применить одну из реализаций алгоритма градиентного бустинга для идентификации программного обеспечения, подобрав оптимальные параметры, сравнить результаты применения данного алгоритма идентификации программ с результатами этого же метода решения данной задачи, но с использованием других видов входных данных.

Основные этапы исследования:

1) Выбор готовой реализации алгоритма градиентного бустинга для исследования.
2) Создание сигнатур рассматриваемых программ:
2.1) изучение сигнатур, которые использовались другими исследователями для распознавания программ с помощью алгоритма градиентного бустинга CatBoost;
2.2) преобразование сигнатур в формат, который можно подать на вход алгоритму.
3) Исследование выбранной реализации алгоритма (основные параметры, системные требования, возможный формат входных данных для обучения и тестирования, формат получаемого результата).
4) Подбор параметров для обучаемой модели.
5) Обучение модели, проведение тестирования и обработка результатов для семи наиболее информативных ассемблерных команд. Анализ полученных результатов. Если результаты неудовлетворительные, возвращение к пункту 4.
6) Составление сводной таблицы результатов. Сравнение всех результатов.

Промежуточные результаты:

Выбрана следующая реализация алгоритма градиентного бустинга: CatBoost. Сигнатуры были преобразованы в шестнадцатеричный вид. Основные параметры алгоритма градиентного бустинга: количество итераций – максимальное число деревьев, которое будет построено при решении задачи машинного обучения; скорость обучения – скорость обучения, используемая для уменьшения шага градиентного спуска; глубина исследования – количество уровней построенного дерева, которые изучаются для классификации программ. Оптимальные параметры для CatBoost: количество итераций – 1000, скорость обучения – 0,18, глубина исследования – 2.

Основной результат:

Получен способ идентификации программного обеспечения с помощью CatBoost, точность идентификации – 77,24%, что является более низким по сравнению с тем же CatBoost с использованием дизассемблированных сигнатур программ (86%), но выше, чем для каждого байта тех же сигнатур в отдельности (68,29%).

Автор: _____/(В.С. Багно)

Научный руководитель: _____/(И.Е. Кривцова)

Декан: _____/(Д.А. Заколдаев)