1. 004.02
2. Применение алгоритма градиентного бустинга в задаче идентификации программного обеспечения
3. В.В. Малков
4. И.Е. Кривцова
5. Основные части тезиса

Краткое введение:

Объект – методы идентификации программного обеспечения, предмет – алгоритм градиентного бустинга. В сфере информационной безопасности сегодня существует множество областей применения машинного обучения. Задача идентификации различных программ является одной из актуальных проблем. В основном, на практике пытаются найти запрещенные и нелицензионные программные средства. В данной работе предложен способ идентификации по сигнатурам исходного кода, преобразованного в ассемблерный, с помощью алгоритма градиентного бустинга, который является одним из алгоритмов машинного обучения.

Цели работы:

Изучить реализации алгоритма градиентного бустинга, применить одну из реализаций алгоритма градиентного бустинга для идентификации программного обеспечения, сравнить результаты применения данного алгоритма идентификации программ с результатами других методов решения данной задачи.

Основные этапы исследования:

1) Выбор готовой реализации алгоритма градиентного бустинга для исследования.
2) Создание сигнатур рассматриваемых программ:
2.1) изучение сигнатур от предыдущих исследователей, которые использовались для распознавания программ с помощью алгоритма градиентного бустинга CatBoost;
2.2) преобразование сигнатур в формат или форматы, который можно подать на вход алгоритмам, выбранным для исследования.
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 десяти наиболее информативных ассемблерных команд. Анализ полученных результатов, если результаты неудовлетворительные, то возвращаемся к пункту 4.
6) Составление сводной таблицы результатов. Сравнение всех результатов.

Промежуточные результаты:

Выбрана следующая реализация алгоритма градиентного бустинга: LightGBM. Сигнатуры были преобразованы в формат csv. 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 LightGBM: количество итераций – 1000, скорость обучения – 0.3, глубина исследования – 7.

Основной результат:

Получен способ идентификации программного обеспечения с помощью LightGBM, точность идентификации – 86,9%. Что является более низким по сравнению с Catboost(89%), но выше результатов с применением статистического критерия(62%).

Автор: _____/(В.В. Малков)

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

Декан: _____/(Д.А. Заколдаев)