УДК 004.8

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ ВРЕДОНОСНЫХ URL

Рыжова Валерия Сергеевна (ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»)

Научный руководитель – доцент,кандидат технических наук Штенников Дмитрий Геннадьевич

(ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»)

Определение вредоносного URL с помощью машинного обучения. Нахождение оптимального решения при использовании разных методов классификации.

Введение. Одна из наиболее быстрорастущих проблем в сфере безопасности – программы-вымогатели, часто распространяющиеся через рассылку вредоносных ссылок. Далеко не все присылаемые нам ссылки, как друзьями, так и незнакомцами, являются безопасными. В современном мире есть множество средств для коммуникации: электронная почта, SMS, социальные сети (например, Facebook и Twitter), приложения для совместной работы и так далее. В большинстве случаев ссылки являются безопасными, но иногда нет. Однако, даже единичные инциденты могут доставить массу неприятностей и даже привести к потере большого количества информации или денег. Остальные виды опасностей и фишинговые сайты также представляют серьезную угрозу. На сегодняшний день определение вредоносных URL основывается на сравнении с базой данных злокачественных ссылок и выдают результат на основе сравнения полученных и имеющихся данных.

Основная часть. Данная работа направлена на сравнительное исследование трех бинарных классификаторов, таких как логическая регрессия, случайный лес и дерево решений, направленных на решения задачи определения опасных интернет-страниц.

Программная реализация моделей была описана на языке Python с использованием таких библиотек как Numpy, pandas для обработки данных и математических вычислений, matplotlib.pyplot и seaborn для визуализации необходимых данных.Seaborn является хорошим дополнением для matplotlib и тесно интегрируется со структурами данных pandas.Так же была использована библиотека tld для получения верхнего уровня домена из URL ,а также sklearn для реализации трех моделей классификации использованных в работе. Для случайного леса была использована sklearn.ensemble , для дерева решений sklearn.tree , а для логической регрессии sklearn.linear_model.

Для обучения модели была выбрана размеченная выборка на открытом ресурсе Kaggle.Этот набор данных содержит записи вредоносных и доброкачественных URL-адресов, которые могут быть использованы для анализа или построения классификаторов, что отлично подходит для нашего исследования.

Для разработки классификатора были выделены важные признаки для определения вредоносных URL, по которым мы и строили нашу модель.

Для того чтобы оценить работу модели были использованы такое метрики ,как confusion matrix,accuracy_score,recall,precision.

Выводы.

Данное исследование нацелено на сравнительный анализ классификаторов для решения предложенной проблемы. В результате все методы показали почти идеальный результат классификации, что может быть вызвано дисбалансом классов в выборке. Решение данной проблемы в будущем поможет повысить точность в дальнейших исследованиях на более широких выборках.

Рыжова В.С. (автор) Подпись

Штенников Д.Г.(научный руководитель) Подпись