Подбор оптимальных параметров для методов машинного обучения при обнаружении вредоносных запросов к веб-приложениям

А. О. Болгов
Кафедра Автоматики и Телемеханики
Пермский национальный исследовательский политехнический университет
aleksynderbolgov@gmail.com

А. О. Каменских
Кафедра Автоматики и Телемеханики
Пермский национальный исследовательский политехнический университет
antoshkinoinfo@yandex.ru

***Аннотация*. Межсетевые экраны по-прежнему являются одной из ключевых технологий защиты веб-приложений от современных киберугроз. Стратегия всесторонней защиты начинается с изоляции с помощью брандмауэров и продолжается другими системами защиты, такими как системы обнаружения вторжений. Проблема с брандмауэрами — ложные срабатывания, которые можно устранить с помощью дополнительных инструментов фильтрации. Использование методов машинного обучения – одно из возможных направлений развития систем защиты. В статье представлен выбор оптимальных параметров для нескольких методов классификации, используемых в машинном обучении. Для этой задачи используется набор обучающих данных с распространенными атаками на веб-приложениях.**

Ключевые слова: безопасность веб-приложений, машинное обучение, методы классификации, логистическая регрессия, классификатор дерева решений

I. Введение

С каждым годом разрабатывается и внедряется все больше веб-приложений. Только доменных имен за 2020 год в общем счета стало 366.3 миллиона [1]. Так как веб-приложения обрабатывают важную бизнес информацию, персональные данные и другие конфиденциальные данные, то вместе с ростом числа веб-приложений растет и количество кибер-атак. По данным Positive Technologies в год фиксируется около 230 тысяч кибер-атак на веб-приложения [2], их реальное число может быть значительно выше. В основе таких атак могут находиться стандартные веб-уязвимости.

К наиболее популярным уязвимостям относятся SQL Injection, Path Traversal и XSS. Так в 53 % проанализированных веб-приложений были обнаружены возможности для реализации атаки межсайтового скриптинга [3].

Такое количество атак неизбежно приводит к усложнению их обнаружения. В большинстве случаев, чтобы обнаружить сложные атаки требуется полноценный программный комплекс вроде IDS, IPS или SIEM, которым также требуется какое-то время на интеллектуальную фильтрацию входящего трафика, и цена таких решений часто высока [4]. Следовательно, в качестве альтернативы дорогостоящим программным решениям можно применить комбинацию из нескольких недорогих программных продуктов при совместном использовании с ними машинного обучения.

В статье приведены оптимальные значения параметров для алгоритмов машинного обучения:

- метод K-ближайших соседей;
- логистическая регрессия;
- дерево решений.

В качестве данных для обучения исследуемых выше моделей используется github репозиторий [5]

II. Методика обнаружение атак при помощи машинного обучения

Усложнение кибер-угроз и повышение их числа приводит к необходимости внедрения концептуально новых методов защиты конфиденциальной информации. Например, в условиях постоянной изменчивости сигнатур конкретны х атак, довольно эффективными будут механизмы защиты, позволяющие находить закономерности в данных с низкой корреляцией. К таким механизмам можно отнести машинное обучение [6]. Сама методика обнаружения атаки при помощи машинного обучения заключается не только в обнаружении закономерностей в данных, но и таких закономерностях, которые могут сигнализировать об атаке.

Для обработки с помощью модели машинного обучения данные должны быть представлено в закодированном виде [7]. Стандартно ML системы работают с весами (числами). Следовательно, перед поиском в данных признаков атаки необходимо эти данные преобразовать в веса, с которыми уже смогут работать модели машинного обучения. Функциональная схема процесса работы машинного обучения по распознаванию атаки представлена на рис. 1.

Рис. 1. Процесс распознавания атаки при помощи машинного обучения

Таким образом, методика распознавания кибер-атак при помощи машинного обучения заключается в замене сигнатурных методов распознавания атак методами машинного обучения, а также добавлением пред- и пост-обработки данных.

III. Анализ модели K-ближайших соседей

Метод K-ближайших соседей является одним из самых простых алгоритмов машинного обучения, применяемых для классификации. Основой метода является сохранение всего обучающего набора данных и дальнейшее сопоставление каждого входного набора признаков с K-ближайшими к нему [8]. К преимуществам метода можно отнести простоту реализации, наглядность принципа работы и отсутствия необходимости сложного предварительного обучения. Недостатками алгоритма являются высокая сложность каждого прогноза (буквально необходимо сопоставлять каждый тестовый вектор данных со множеством всех сохраненных данных). Из первого недостатка также вытекает и второй – проблема размерности. В отличие от большинства других алгоритмов машинного обучения, производительность алгоритма K-ближайших соседей сильно зависит от размера обучающей выборки. Чем больше выборка, тем дольше будет вычисляться прогноз [9].

В рамках статьи осуществляется подбор оптимального значения K-соседей в алгоритме K-ближайших соседей. Для каждого объема данных обучение и валидация на тестовой выборке происходили 5 раз, чтобы уменьшить влияние потенциальных погрешностей при моделировании.

На рис. 2 представлен график зависимости точности на тестовой выборке и количества соседей для данных объемом 100/1000, 1000/10000, 4000/40000, 8000/80000.

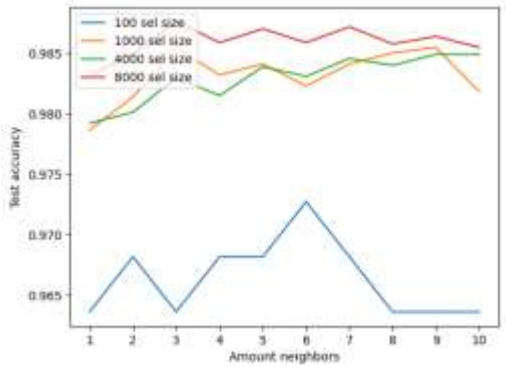

Рис. 2. Графики зависимости точности на тестовой выборке и количестве K-соседей

По графику можно сделать вывод, что наиболее высокие значения точности на тестовой выборке достигаются при использовании значений параметра K-соседей от 7 до 9, за исключением данных по первому объему данных (100/1000). Это можно объяснить недостаточным размером обучающей выборки. Модель при обучении на 1000 примерах не смогла найти закономерность в многообразии данных и не научилась обобщать их на новые данные при использовании большого значения параметра K-соседей. Также можно наблюдать общее повышение точности классификации на тестовой выборке с увеличением объема данных для обучения, но ввиду особенности работы модели K-ближайших соседей увеличение объема данных для обучения многократно увеличивает требуемое время на обучение и классификацию на тестовой выборке.

IV. Анализ модели логистической регрессии

Логистическая регрессия является одним из статистических методов анализа данных. В отличие от линейной регрессии, в модели логистической регрессии вместо предсказания числовой переменной производится расчет вероятности принадлежности числового значения к тому или иному классу [10–11].

В рамках статьи исследуются такие параметры как:

1. вид регуляризации. Сравниваются между собой регуляризации вида L1, L2, регуляризации гибкой сети и полное отсутствие регуляризации[12–13];
2. методы минимизации ошибки. Сравниваются между собой метод Ньютона, LBFGS-метод, SAG-метод, SAGA-метод[14–16].

На рис. 2–5 представлены зависимости точности измерений от размера выборки, используемого метода минимизации ошибки и вида регуляризации.

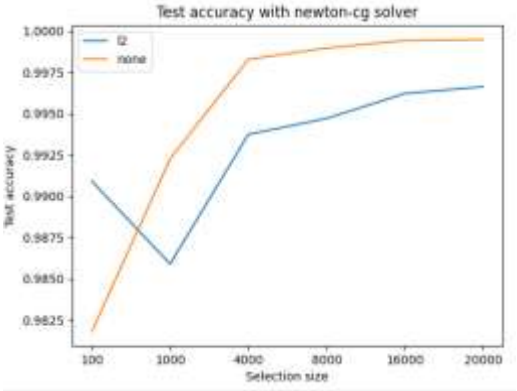

Рис. 3. Зависимость точности классификации от вида регуляризации и объема данных при использовании метода Ньютона

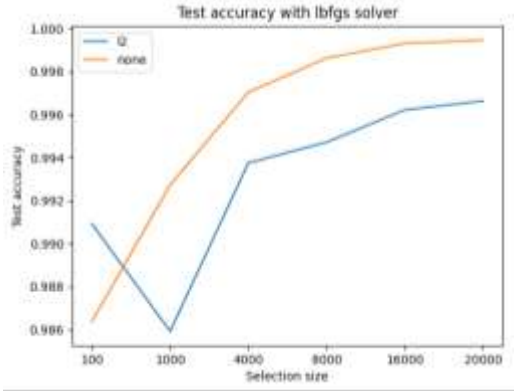

Рис. 4. Зависимость точности классификации от вида регуляризации и объема данных при использовании LBFGS метода

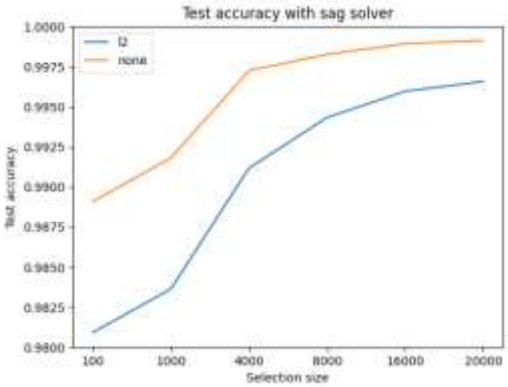

Рис. 5. Зависимость точности классификации от вида регуляризации и объема данных при использовании SAG метода

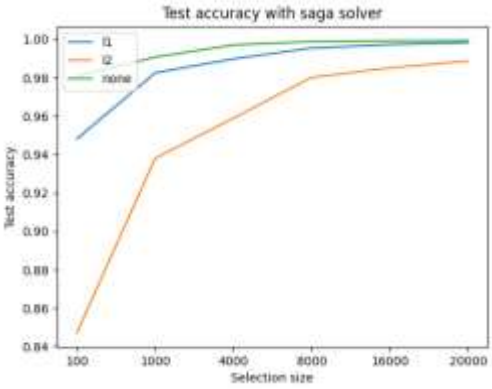

Рис. 6. Зависимость точности классификации от вида регуляризации и объема данных при использовании SAGA метода

По графикам и данным в таблице, можно заключить модель логистической регрессии довольно точно классифицирует данные, принадлежащие тестовой выборке и без регуляризации на всех объемах данных, за исключением 100/1000 запросов. Это может говорить о том, что на приведенных данных для обнаружения вредоносных запросов нет необходимости использовать регуляризацию для исключения переобучения вне зависимости от типа метода минимизации ошибки при обучении.

V. Анализ метода дерево классификации

Под деревом решений в машинном обучении подразумевается древовидная иерархическая структура, состоящая из корня, узлов, листьев решений и связей между ними. В основе каждого узла находится условие формата “Если то …”, это условие разделяет множество примеров S на два подмножества, по определенным признакам обучающего множества. Несмотря на простоту, лежащую в основе этого метода, он достаточно эффективен для решения задач классификации и регрессии [17].

В рамках статьи исследуются следующие параметры модели дерева классификации:

1. способ разделения множества на подмножества. Сравнивается случайное разделение (Random) с разделением по критерию лучшей информативности признака примера (Best);
2. критерий оценки качества разбиения. Сравнивается критерий Джини (Gini) с понижением энтропии (Entropy) [18].

На рис. 7–12 представлены графики зависимости точности классификации от глубины дерева, способа разделения множества и критерия оценки разделения.

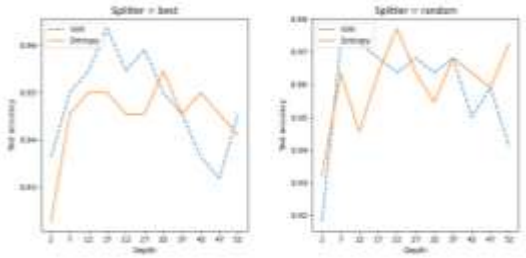

Рис. 7. Зависимость точности классификации от глубины дерева, способа разделения множества и критерия оценки разделения для объема данных 100/1000

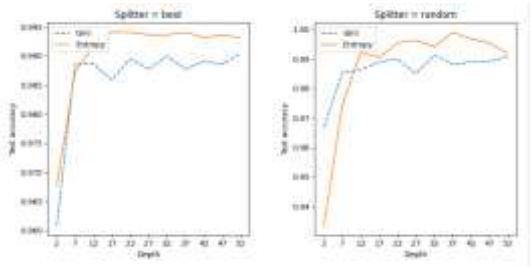

Рис. 8. Зависимость точности классификации от глубины дерева, способа разделения множества и критерия оценки разделения для объема данных 1000/10000

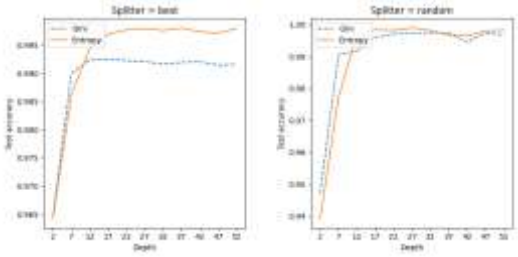

Рис. 9. Зависимость точности классификации от глубины дерева, способа разделения множества и критерия оценки разделения для объема данных 4000/40000

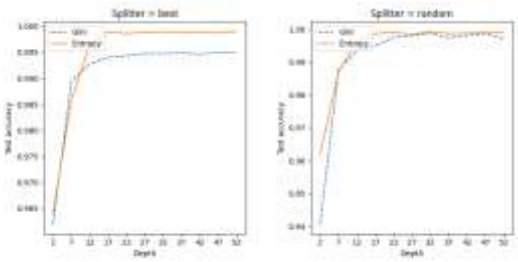

Рис. 10.Зависимость точности классификации от глубины дерева, способа разделения множества и критерия оценки разделения для объема данных 8000/80000

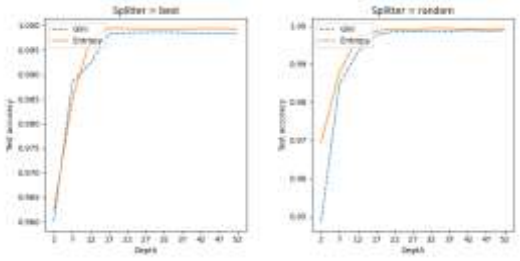

Рис. 11.Зависимость точности классификации от глубины дерева, способа разделения множества и критерия оценки разделения для объема данных 16000/160000

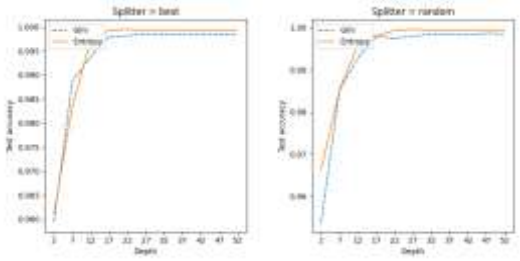

Рис. 12.Зависимость точности классификации от глубины дерева, способа разделения множества и критерия оценки разделения для объема данных 20000/200000

По графикам можно заключить, что с увеличением объема данных для обучения различие между критериями оценки разбиения множества уменьшается, при достижении объема данных 16000/160000 разница между критериями оценки разбиения минимальна. На малом объеме данных предпочтительней использовать критерий оценки Gini.

VI. Анализ применимости моделей распознавания вредоносного трафика

Исходя из принципа работы моделей на основе машинного обучения, представленных в статье, возможно несколько вариантов их интеграции в существующий комплекс средств информационной безопасности. Первый из них – замена существующего межсетевого экрана одной из моделей, либо ее развертывание с аналогичной целью. Второй – параллельное включение одной из моделей с межсетевым экраном.

С одной стороны, первый вариант интеграции позволяет использовать в качестве межсетевого экрана самописную модель распознавания на основе машинного обучения. С другой стороны, разработанная модель обладает меньшим количеством встроенного функционала, в отличие от уже готовых решений в этой области [19].

Второй вариант, когда модель обнаружения вредоносного трафика подключается параллельно с межсетевым экраном. Схематично это включение представлено на рис. 13.

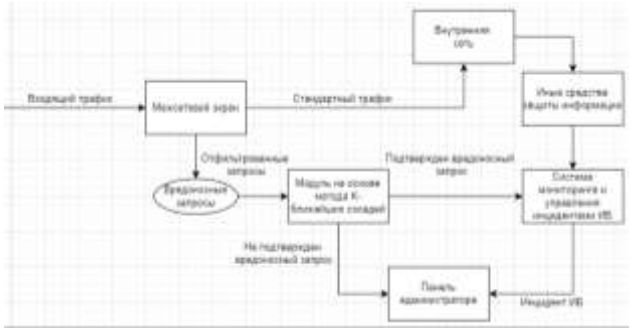

Рис. 13.Схема интеграции модели распознавания вредоносного трафика

На вход модели обнаружения вредоносных запросов подается информация о таких запросах с выхода межсетевого экрана. Далее в случае подтверждения угрозы от таких запросов, модель может уведомить об это систему мониторинга и управления инцидентами информационной безопасности. В случае, если угроза от такого запроса не подтвердится, то модель может уведомить администратора информационной безопасности о потенциально ложном срабатывании.

VII. Заключение

По результатам моделирования, можно сделать следующие выводы для каждой из исследуемых моделей машинного обучения.

Для модели K-ближайших соседей оптимальным параметром K-соседей для обучения является интервал из 7–9 соседей. Уменьшение числа соседей приводит к усложнению модели машинного обучения, и следовательно ей не хватает того объема и многообразия данных, что был представлен в выборке. Увеличение числа соседей приводит к упрощению модели, из-за этого, вероятно, обученная модель будет недостаточно точно обобщать закономерности в обучающих данных на новые (тестовые) данные.

Для модели логистической регрессии сравнивались виды регуляризации и методы минимизации ошибки. По результатам обучения установлено, что отсутствие регуляризации положительно сказывается на точности классификации модели, вследствие этого можно прийти к двум выводам:

1. модель попросту не успевала переобучиться, так как данных было недостаточно;
2. соотношение объема данных для обучения и их многообразие было оптимальным для обучения.

При этом, если учесть, что точность на тестовых данных при отсутствии регуляризации также была достаточно высокой, что видно по графикам. Следовательно, при обучении модели для обнаружения вредоносных запросов на представленных данных предпочтительнее не использовать регуляризацию.

Для модели дерево классификации можно сделать вывод о том, что разница при использовании разных критериев оценки качества разбиения множества признаков, уменьшается с увеличением обучающей выборки, тем не менее, при обучении на небольшой выборке следует использовать критерий Gini. Разница между видом разбиения также незначительна, но тем не

менее, разбиение с учетом информативности признака оказывается предпочтительней.

Список литературы

[1] The Domain Name Industry Brief. [Электронный ресурс]. - URL: https://www.verisign.com/en_US/domain-names/dnib/index.xhtml (Дата обращения 29.11.2021)

[2] Атаки на веб-приложения: итоги 2018 года. [Электронный ресурс]. - URL: https://www.ptsecurity.com/ru-ru/research/analytics/web-application-attacks-2019/ (Дата обращения 08.12.2021)

[3] Переобучение. [Электронный ресурс]. -URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9F%D0%B5%D1%80%D0%B5%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5 (Дата обращения 10.12.2021)

[4] Обзор мирового и Российского рынка SIEM-систем. [Электронный ресурс]. -URL: https://www.anti-malware.ru/analytics/Market_Analysis/overview-global-and-russian-market-siem (Дата обращения 13.12.2021)

[5] Fwaf-Machine-Learning-driven-Web-Application-Firewall. -URL: https://github.com/faizann24/Fwaf-Machine-Learning-driven-Web-Application-Firewall (Дата обращения 20.11.2021)

[6] Amal Saha, Sugata Sanyal. Application Layer Intrusion Detection with Combination of Explicit-Rule- Based and Machine Learning Algorithms and Deployment in Cyber- Defence Program. arXiv, Cornell University. Computer Science, Cryptography and Security. November 2014

[7] Ozlem Yavanoglu, Murat Aydos. A review on cyber security datasets for machine learning algorithms. IEEE International Conference on Big Data (Big Data). Dec 2017.

[8] J.Laaksonen, E.Oja. Classification with learning K-nearest neighbors. Helsinki University of Technology Laboratory of Computer and Information Science.

[9] Sarah Jane Delany, Padraig Cunningham. K-Nearest Neighbor Classifier. ACM Computing Surveys, April 2007.

[10] Maher Maalouf. Logistic regression in data analysis: an overview. International Journal of Data Analysis Techniques and Strategies, July 18 2011.

[11] Stephan Dreiseitl, Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, October 2002.

[12] Jigyasu Joshi, Shweta Saxena. Regression analysis in data science. Journal of Analysis and Computation, December 2020.

[13] Diego Vidaurre, Concha Bielza, Pedro Larranaga. A Survey of L1 Regression. International Statistical Review, December 2013.

[14] Liu Yang, Yanping Chen, Xiaojiao Tong, Chunlin Deng. A new smoothing Newton method for solving constrained nonlinear equations. Applied Mathematics and Computation, August 2011.

[15] Mehiddin Al-Baaali, Emilio Spedicato, Francesca Maggioni. Broyden's quasi-Newton methods for a nonlinear system of equations and unconstrained optimization: a review and open problems. Optimization Methods and Software, November 2013.

[16] Sebastian Ruder. An overview of gradient descent optimization algorithms. Insight Centre for Data Analytics, Jun 2017.

[17] Chun-ChiehYang, Shiv O Prasher, Peter Enright, Chandra Madramootoo, Magdalena Burgess, Pradeep K Goel, IanCallum. Application of decision tree technology for image classification using remote sensing data. Agricultural Systems, June 2003.

[18] Laura Elena Raileanu, Kilian Stoffel. Theoretical Comparison between the Gini Index and Information Gain Criteria. Annals of Mathematics and Artificial Intelligence, 2004.

[19] Обзор рынка защиты веб-приложений(WAF) в России и в мире [Электронный ресурс]. -URL: https://www.anti-malware.ru/reviews/web_application_firewall_market_overview_russia (Дата обращения: 15.12.20)