DOI 10.37882/2223-2966.2023.12.18

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ КИБЕРАТАК НА ОСНОВЕ СОЦИАЛЬНОЙ ИНЖЕНЕРИИ

APPLICATION OF MACHINE LEARNING METHODS FOR DETECTING CYBER ATTACKS BASED ON SOCIAL ENGINEERING

N. Kotikov
A. Rusakov
V. Filatov

Summary. This article presents a study on the development of a tool for proactive analysis and classification of potentially malicious web pages. The research materials contain an overview of the existing means of identifying illegitimate resources, justified by the relevance of the study. Also presented are: scope of application, subject of research and potential limitations of software implementation. The main tasks are formulated, algorithms and methods that need to be taken into account during development are defined. The article concludes that an integrated approach to ensuring infrastructure protection, taking into account multivector analysis, is quite in demand both theoretically and practically.

Keywords: classification of web pages, phishing, information security.

Котиков Никита Михайлович
Российский Технологический Университет МИРЭА, Москва
KotikovNik@yandex.ru
Русаков Алексей Михайлович
Старший преподаватель, Российский Технологический Университет МИРЭА, Москва
rusal@bk.ru
Филатов Вячеслав Валерьевич
Кандидат технических наук, доцент, Российский Технологический Университет МИРЭА, Москва
filv@mail.ru

Аннотация. В данной статье представлено исследование по разработке инструмента проактивного анализа и классификации потенциально вредоносных веб-страниц. В материалах исследования содержится обзор существующих средств выявления нелегитимных ресурсов, обоснованный актуальностью исследования. Также представлены: область применения, предмет исследования и потенциальные ограничения программной реализации. Сформулированы основные задачи, определены алгоритмы и методы, которые необходимо учесть при разработке. В статье производится вывод, что комплексный подход к обеспечению защиты инфраструктуры с учетом многовекторного анализа является достаточно востребованным как в теоретическом, таки в практическом плане.

Ключевые слова: классификация веб-страниц, фишинг, информационная безопасность.

В современном мире компьютерных технологий представители малого и среднего бизнеса, крупных корпораций и даже рядовых пользователей с личными устройствами для доступа в сеть интернет так или иначе знакомы с такими понятиями как защита информации, хакеры, утечки информации. В рамках чтения публичных средств массовой информации мы часто видим все новые и новые статьи о хакерских атаках и их последствиях — все это дополнительно, несмотря на требования регуляторов по информационной безопасности, добавляет понимания и осознанности специалистам, что любую информационную систему (далее — ИС) необходимо обеспечить защитой от различного рода векторов нелегитимного воздействия. Можно наблюдать, как, построив саму ИС, выстроив процессы взаимодействия и реагирования на возникающие инциденты, организации начинают чувствовать себя в относительной безопасности. Но зачастую это не так, ведь представители компаний не учитывают один важный факт — пользователями ИС являются обычные люди. Несомненно, использование криптографии, средств межсетевого экранирования, контроля доступа и других решений в разрезе построения комплексной системы обеспечения информационной безопасности на предприятии обеспечивает надежный уровень эшелонированной защиты. Но именно взаимодействие человека с системой порой несет наиболее неявные и критичные риски.

Человеческий фактор

Информация представляет собой один из наиболее ценных активов для компании, который может содержать в себе тайну или иные конфиденциальные данные, при определенных условиях позволяющие укрепить положение на рынке или избежать непредвиденных потерь. Именно поэтому так много внимания следует уделять ее защите. С любой ИС взаимодействуют пользователи — люди, что неизбежно влечет возникновение влияния человеческого фактора на все без исключения процессы организации.

Человеческий фактор — термин, которым описывают неконтролируемое влияние на принятие решений личностями в ситуации, обусловленной психологической неустойчивостью отдельной штатной единицы в условиях потенциального риска нарушения информационной безопасности.

Соответственно, человеческий фактор определяет две категории действий личности.

К умышленным действиям относятся: намеренная кража, предоставление конфиденциальной и иной, имеющей ценность для организации, информации, ее искажение, модификация и уничтожение, иными словами, нарушение любого из свойств информации.

Соответственно, к неумышленным можно отнести утрату, уничтожение носителей или самой информации по неосторожности, когда человек допускает нарушение информационной безопасности случайно. Также к этой категории относится раскрытие информации под влиянием методов социальной инженерии.

Социальная инженерия и ее виды

Социальная инженерия на данный момент представляет собой совокупность различных техник несанкционированного доступа к защищаемой информации без использования технических средств для воздействия на целевую инфраструктуру. Основным методом является использование психологических приемов, направленных непосредственно на эксплуатацию человеческого фактора. Злоумышленник, используют все новые и новые техники общения с жертвами с целью добиться доверия и получить интересующую информацию. Социальная инженерия в первую очередь направлена на доверчивых и невнимательных сотрудников. На данный момент методы психологической манипуляции являются полноценным разделом в информационной безопасности и изучаются на уровне государства.

Можно выделить несколько наиболее актуальных методов.

Фишинг (англ. phishing произв. fishing — рыбалка) — метод, обусловленный рассылкой сообщений с исполь-

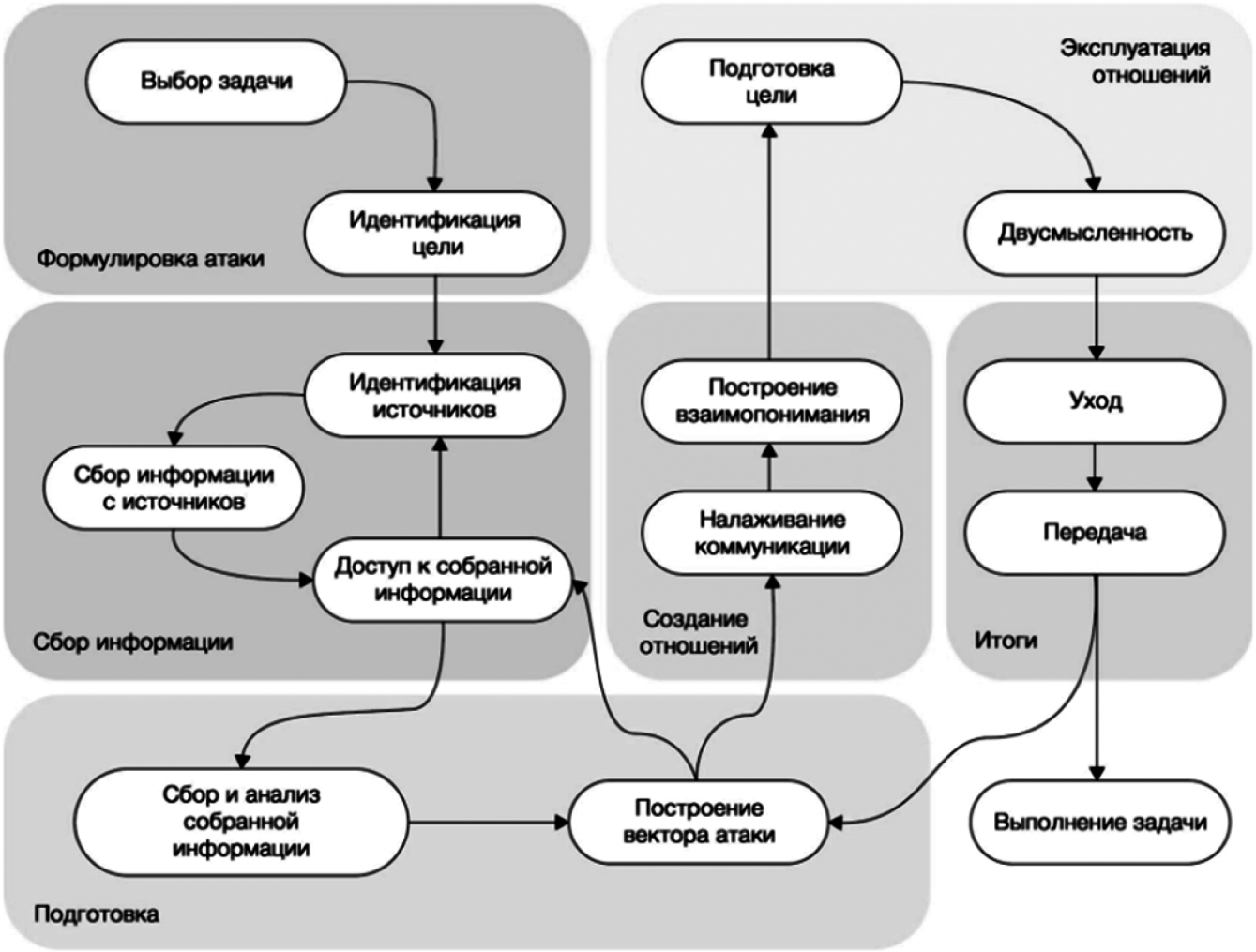

Рис. 1. Схема взаимодействия

зованием почтовых клиентов или мессенджеров, внутри которых содержатся ссылки и qr-коды, ведущие на зловредные ресурсы, или вредоносные файлы под видом обычного документа. Такие рассылки обычно производятся под видом руководства или сотрудников компаний для еще большего введения в заблуждение.

Фарминг (англ. farming — культивация в сельском хозяйстве) — метод, основным вектором которого является доставка пользователям вредоносного файла, перенаправляющего трафик на поддельные копии реально существующих ресурсов.

Вишинг (англ. voice fishing — голосовой фишинг, сокращённо «vishing») — метод, основанный на выполнении голосовых вызовов от имени компаний или силовых ведомств. Яркий пример — звонки от «службы безопасности Сбербанка» или «сотрудников Госуслуг».

Байтинг (англ. bait — наживка) — метод, основанный на размещении приманки, которая заставит цель инициировать взаимодействие. К примеру, размещение съемного носителя с вредоносной нагрузкой.

Зачастую хорошо продуманные целевые атаки содержат в себе совокупность разных методов и тактик, позволяющих получить необходимую информацию и развить воздействие.

Используемые классы решений по защите информации

На данный момент большинство организаций защищены от воздействия с использованием основных методов социальной инженерии.

Для защиты конечных точек, контроля подключений съемных носителей и загрузки файлов с целью проверки на наличие вредоносных файлов компании используют антивирусные средства защиты (AV) и средства защиты конечных точек (EDR).

Активно внедряются классы решений, предназначенных для предотвращения передачи конфиденциальной информации по каналам связи, контроля действий пользователей на устройстве в реальном времени (DLP).

Для предотвращения стандартных фишинговых и спам — рассылок организации используют почтовые шлюзы — (Mail Gateway), средства защиты, которые анализируют заголовки поступающей электронной корреспонденции, домен отправителя и содержимое. Данный класс решений иногда сочетает в себе функционал песочницы (SanBox) для анализа поступающих файлов.

Стоит отметить, что даже совокупность всех перечисленных средств защиты в комплексе (комплексная система обеспечения информационной безопасности — КСОИБ) и грамотно выстроенные процессы информационной безопасности не позволяют гарантировать высокий уровень защиты пользователей и организации в целом от хорошо продуманных целевых атак с использованием социальной инженерии, так как необходима не просто защита от файловых угроз, контроль конечных точек и почты, но и проактивный анализ содержимого интернет — ресурсов при взаимодействии с браузером пользователей.

Обзор и анализ существующих алгоритмов машинного обучения

Машинное обучение — довольно крупный раздел, являющийся одной из форм искусственного интеллекта, изучающий методы построения алгоритмов, основным принципом обучения которых является обучение на основании решений множества аналогичных задач.

Байесовский классификатор.

Центральным инструментом машинного обучения считается теорема Байеса, основной принцип которой обусловлен оценкой вероятности не только происхождения события, но и его достоверности:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{P(\theta)P(D|\theta)}{\int_{\theta\in\Theta} P(D|\theta)P(\theta)d\theta} \quad (1)$$

где θ — параметры модели,

D — данные объекта исследования,

$P(\theta)$ — априорная вероятность (prior probability),

$P(D \mid \theta)$ — правдоподобие (likelihood),

$P(\theta|D)$ — апостериорная вероятность (posterior probability),

$P(D) = \int P(D|\theta)P(\theta)d\theta$ — вероятность данных (evidence).

Задачей машинного обучения является поиск и/или максимизация распределения апостериорной вероятности $P(\theta|D)$ в целях определения наиболее подходящих к набору данных параметров для воспроизведения новых предсказаний.

Главный принцип классификатора заключается в вычислении функции правдоподобия для каждого объекта класса, по которой, в свою очередь, вычисляются апостериорные вероятности классов. Для этого необходимо определить плотность распределения каждого из классов. Данный метод не труден в реализации, но требует обучающей выборки, которая сильно влияет на точность классификатора своим объемом.

Алгоритм k-ближайших соседей

Данный метод является метрическим классификатором, обусловленным оценкой сходства объектов. Алго-

ритм причисляет исследуемый объект к классу, которому принадлежат ближайшие его соседи. Для реализации алгоритма используется расстояние Минковского:

$$\rho(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (2)$$

где n — количество объектов в наборе данных,

p — параметр Минковского.

Метод ближайших соседей несложен в реализации с любым количеством параметров в классификации, но увеличение объема обучающей выборки негативно влияет на алгоритм, помимо этого он требует существенных вычислительных мощностей и обладает не самое лучшей точностью классификации.

Логистическая регрессия

Регрессии применяются для моделирования отношений между переменными, а также дальнейшего анализа с целью определения влияния этих переменных на итоговый результат. Регрессия является зависимостью математического ожидания случайного значения от определенного количества других величин, имеет общий вид $E(y \mid x) = f(x)$. Основной задачей выступает поиск функции, которая наиболее точно описывает текущую зависимость. Одним из алгоритмов обучения для методов машинного обучения является линейная регрессия, которая обусловлена моделью зависимости линейного вида:

$$y(x,\omega) = \omega_0 + \sum_{j=1}^{p} x_j w_j = x^T w \quad (3)$$

Метод логистической регрессии широко применяется для решения линейных задач (неприменимо для нелинейных), прост в реализации, не требует большого количества вычислительных мощностей, но является не самым мощным и может быть превзойден другими алгоритмами.

Искусственные нейронные сети

Данный метод является математической моделью, построенной по принципу нейронных связей живых организмов, где каждый нейрон обладает связью с другими слоями соседей, а их эффективность повышается по мере обучения:

$$y(x,w) = f\left(\sum_{j=1}^{N} w_j \phi_j(x) \right) \quad (4)$$

где f — нелинейная функция активации,

w — вектор весов,

ϕ — нелинейные базисные функции.

Обучение применяет принцип «обратного распространения ошибки», при котором разница эталонного и полученного значения передается между слоями в соответствии с коэффициентами связанности между нейронами

Сравнение программного обеспечения в области классификации фишинговых сайтов

Существует большое количество исследований на тему использования методов искусственного интеллекта для анализа легитимности интернет — ресурсов.

Для их сравнения необходимо получить среднеарифметическое значение качества модели по каждому из методов, учитывая частоту применения. Для простоты работы будут обозначены от W1 до W14 (см. табл. 1).

На основе сравнительного анализа можно сделать вывод, что наиболее востребованным является ансамбль случайного леса — при малых затратах наибольшая точность классификатора. Но наилучшего результата классификации позволяет достичь алгоритм градиентного бустинга XGBoost на основе деревьев решений.

Критерии классификации интернет-ресурсов

В качестве основного критерия определения подлинности веб-страниц выступают компоненты URL.

<схема> : [// [<логин> [: <пароль>] @] <хост> [: <порт>]] [/ <URL адрес>] [? <параметры>] [# <якорь>]

Рис. 2. Структура URL

<схема> — схема обращения к ресурсу, т.е. сетевой протокол;

<логин> — имя пользователя, используемое для доступа к ресурсу;

<пароль> — пароль пользователя;

<хост> — полное доменное имя хоста в системе DNS или IP-адрес;

<порт> — порт хоста для подключения;

<URL-адрес> — унифицированный указатель ресурса;

<параметры> — строка GET-запроса с передаваемыми параметрами на сервер (символ «?» — начинало запроса, символ «&» — разделитель параметров);

<якорь> — идентификатор «якоря» с предшествующим символом #.

Таблица 1.

Сравнение применяемых алгоритмов в существующих решениях

Метод	W7	W1	W12	W3	W5	W9	W10	W2	W8	W4	W6	W11	W14	W13	Σ	%
	99,3	98,7	98,0	97,9	97,8	97,6	97,4	96,5	95,3	94,3	93,4	92,7	91,5	83,0		
XGBoost		+													1	98,72
Extra-Tree		+	+												2	98,36
SVM			+												1	98,00
LR			+												1	98,00
kNN			+												1	98,00
Adaboost			+												1	98,00
Gradient Boosting			+												1	98,00
Bagging			+				+								2	97,70
CNN						+									1	97,60
Bayesian								+							1	96,50
DT	+		+		+								+		4	96,64
ANN			+				+		+		+				4	96,04
RF		+	+	+			+			+	+			+	7	94,69

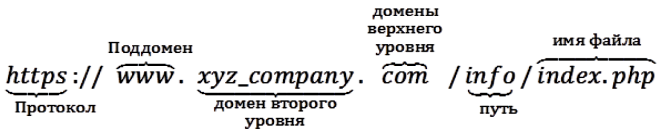

Рис. 3. Общий формат URL

Первым компонентом URL является протокол доступа к ресурсу, второй определяет имя узла или его ip-адрес (разделяется на поддомен, основной и домен верхнего уровня). Далее третьим компонентом является идентификатором конкретного объекта, запрошенного пользователя.

Соответственно, злоумышленники, в целях повышения успешности своих целевых воздействий, используют методы модификации URL, cybersquatting, typosquatting, punycode и т.д.

Для распознавания подлинности интернет-страниц используется комплекс критериев, каждый из которых описан в работах о том, как оценивать фишинговые страницы при помощи методов искусственного интеллекта. Для простоты рассматриваемых работ были использованы обозначения от W1 до W12.

Все представленные ниже критерии были отсортированы по условной величине:

$$Rank = \frac{\sum_{i=1}^{n} r_i * a_i}{n} \quad (5)$$

где n — кол-во моделей;

r_i — точность классификации модели m_i, использующей рассматриваемый критерий;

a_i — наличие рассматриваемого критерия.

Критерии, описанные в рассматриваемых работах, сведены в таблицу 2.

Алгоритм классификации

На старте классификации происходит сбор необходимых данных для классификации, включая перечень необходимых критериев для классификации, обученную модель и перечень компонентов. Далее происходит комплексный компонентно-сигнатурный анализ, результатом которого является положительный либо отрицательный результат проверки подлинности.

Тестирование программного продукта

В ходе эксплуатационных испытаний программного комплекса была проведена серия экспериментов, а решение позволяет производить эффективную оценку подлинности интернет-ресурсов (см. рис. 4).

Заключение

Были рассмотрены и проанализированы методы и алгоритмы реализации программных средств с использованием методологии машинного обучения для

Таблица 2.

Признаки фишинговых сайтов

Критерий	W12	W1	W9	W2	W6	W7	W5	W3	W4	W8	W11	W10	Σ	%	Rank
	99,9	98,7	98,0	98,0	97,6	97,4	95,3	94,3	93,4	92,7	91,5	83,0			
1. Длина URL	+	+		+	+	+	+	+	+	+	+		10	95,9	79,9
2. Наличие IP-адресов в URL	+	+			+	+	+		+	+	+		8	95,8	63,9
3. Наличие @ в URL	+		+		+	+	+	+	+				7	96,6	56,3
4. Рейтинг сайта	+			+	+	+	+		+	+			7	96,3	56,2
5. Возраст домена	+				+	+	+		+	+	+		7	95,4	55,7
6. HTTPS в основном URL		+			+	+	+	+	+				6	96,1	48,1
7. Соотношение внешних ссылок		+			+	+	+		+		+		6	95,7	47,8
8. Кол-во записей домена в DNS	+				+	+	+		+				5	96,7	40,3
9. Срок регистрации домена	+				+	+	+		+				5	96,7	40,3
10. Соотношение общих страниц (распространённых якорных ссылок)		+			+	+	+		+				5	96,5	40,2
11. Наличие перенаправлений с использованием ‘//’ в URL			+		+	+	+		+				5	96,4	40,1
12. Число поддоменов в URL				+	+	+	+		+				5	96,3	40,1
13. Кол-во якорных ссылок					+	+	+		+		+		5	95,1	39,6
14. Наличие всплывающих окон с полями для ввода текста					+	+	+		+		+		5	95,1	39,6
15. Кол-во повторов запроса URL					+	+	+		+		+		5	95,1	39,6
16. Веб-трафик					+	+	+		+		+		5	95,1	39,6
17. Число слешей в URL	+	+	+					+					4	97,7	32,6
18. Использование службы сокращения URL-адресов «TinyURL»			+		+	+	+						4	97,1	32,4
19. Нестандартный порт					+	+	+		+				4	95,9	32,0
20. Отключение события щелчка правой кнопкой мыши					+	+	+		+				4	95,9	32,0
21. Использование Iframe					+	+	+		+				4	95,9	32,0
22. Кол-во перенаправлений					+	+	+		+				4	95,9	32,0
23. Число точек в URL		+	+					+		+			4	95,9	32,0
24. Обработчик серверных форм (SFM)					+	+	+				+		4	95,5	31,8
25. Нулевые ссылки (наличие якорных ссылок)	+	+							+				3	97,4	24,3
26. Валидация TLD		+		+				+					3	97,0	24,2
27. Длина домена		+		+				+					3	97,0	24,2
28. Число имен брендов в URL		+		+				+					3	97,0	24,2
29. Средняя длина слова в URL		+		+				+					3	97,0	24,2
30. Число токенов в URL		+		+				+					3	97,0	24,2
31. Наличие префикса или суффикса, разделенного «-» в домене					+	+	+						3	96,8	24,2
32. Загрузка Favicon с внешнего домена					+	+	+						3	96,8	24,2
33. Кол-во ссылок в тегах <Meta>, <Script>, <Link>					+	+	+						3	96,8	24,2

Критерий	W12	W1	W9	W2	W6	W7	W5	W3	W4	W8	W11	W10	Σ	%	Rank
	99,9	98,7	98,0	98,0	97,6	97,4	95,3	94,3	93,4	92,7	91,5	83,0			
34. Отправка информации по электронной почте					+	+	+						3	96,8	24,2
35. Наличие индекса в Google					+	+	+						3	96,8	24,2
36. Составление статистических отчетов					+	+	+						3	96,8	24,2
37. Конечное состояние SSL					+	+					+		3	95,5	23,9
38. Ненормальный URL					+		+			+			3	95,2	23,8
39. Наличие имен доменов в титульнике	+	+											2	99,4	16,6
40. Наличие имен доменов в авторских правах	+	+											2	99,4	16,6
41. Наличие имен доменов в тексте заголовков	+	+											2	99,4	16,6
42. Кол-во цифр в URL			+	+									2	98,0	16,3
43. Проверка имени хоста по IP-адресу	+							+					2	97,1	16,2
44. Наличие фишинговых слов		+						+					2	96,5	16,1
45. Наличие имен брендов в поддомене		+						+					2	96,5	16,1
46. Число цифр в доменном имени		+						+					2	96,5	16,1
47. Длина имени хоста		+						+					2	96,5	16,1
48. Число дефисов в именах хостов		+						+					2	96,5	16,1
49. Цифры в именах хостов		+						+					2	96,5	16,1
50. Наличие фишинговых слов в URL		+						+					2	96,5	16,1
51. Число цифр в URL		+						+					2	96,5	16,1
52. Длина самого большого слова в URL		+						+					2	96,5	16,1
53. Наличие дефисов в URL			+					+					2	96,1	16,0
54. Количество нижних подчеркиваний в имени хоста			+					+					2	96,1	16,0
55. Наличие вопросительного знака в URL			+					+					2	96,1	16,0
56. Наличие «;» в основном URL			+					+					2	96,1	16,0
57. Изменение статус бара при наведении мыши					+				+				2	95,5	15,9
58. Время между текущим и момента уничтожения домена	+												1	100,0	8,3
59. Наличие заголовка и атрибута ключевого слова	+												1	100,0	8,3
60. Наличие ссылок на текущий домен	+												1	100,0	8,3
61. Средняя длина слова		+											1	98,7	8,2
62. Длина самого большого слова		+											1	98,7	8,2
63. Наличие имен доменов в наименьших терминах TF-IDF		+											1	98,7	8,2
64. Коэффициент сходства объектов (их хэшей) с расстоянием Хэмминга		+											1	98,7	8,2
65. Наличие «=» в основном URL			+										1	98,0	8,2
66. Наличие «+» в основном URL			+										1	98,0	8,2
67. Наличие «:» в основном URL			+										1	98,0	8,2
68. Наличие «~» в основном URL			+										1	98,0	8,2

Критерий	W12	W1	W9	W2	W6	W7	W5	W3	W4	W8	W11	W10	Σ	%	Rank
	99,9	98,7	98,0	98,0	97,6	97,4	95,3	94,3	93,4	92,7	91,5	83,0			
69. Наличие «#» в основном URL			+										1	98,0	8,2
70. Наличие «!» в основном URL			+										1	98,0	8,2
71. Наличие «&» в основном URL			+										1	98,0	8,2
72. Наличие «%» в основном URL			+										1	98,0	8,2
73. Длина самого короткого слова в URL				+									1	98,0	8,2
74. Стандартное отклонение длин слов в необработанном списке слов				+									1	98,0	8,2
75. Количество рассматриваемых слов, обработанных в модуле декомпозиции слова				+									1	98,0	8,2
76. Средняя длина рассматриваемых слов, обработанных в модуле декомпозиции слова				+									1	98,0	8,2
77. Число декомпозированных слов				+									1	98,0	8,2
78. Число ключевых слов в URL				+									1	98,0	8,2
79. Число схожих с ключевыми словами слов				+									1	98,0	8,2
80. Число схожих с ключевыми словами брендов				+									1	98,0	8,2
81. Число рандомно сгенерированных слов				+									1	98,0	8,2
82. Число целевых имен брендов в URL				+									1	98,0	8,2
83. Число целевых ключевых слов в URL				+									1	98,0	8,2
84. Число других слов				+									1	98,0	8,2
85. Имя домена состоит из случайного набора символов				+									1	98,0	8,2
86. Длина поддомена				+									1	98,0	8,2
87. Наличие www. com в доменах или поддоменах				+									1	98,0	8,2
88. Punycode				+									1	98,0	8,2
89. Наличие специальных символов в URL				+									1	98,0	8,2
90. Последовательное повторение символа				+									1	98,0	8,2
91. Кастомизация статус бара							+						1	95,3	7,9
92. Длина самого большого слова в имени хоста								+					1	94,3	7,9
93. Наличие нескольких TLD								+					1	94,3	7,9
94. Наличие «$» в URL								+					1	94,3	7,9
95. Наличие запятой в основном URL								+					1	94,3	7,9
96. Наличие '*' в основном URL								+					1	94,3	7,9
97. Наличие символа OR в основном URL								+					1	94,3	7,9
98. Наличие пробела в основном URL								+					1	94,3	7,9
99. Принадлежность хоста к основным фишинговым доменам									+				1	93,4	7,8
100. Оценка доступности										+			1	92,7	7,7
101. Сравнение коэффициентов сжатия в моделях легитимных и фишинговых сайтов												+	1	83,0	6,9

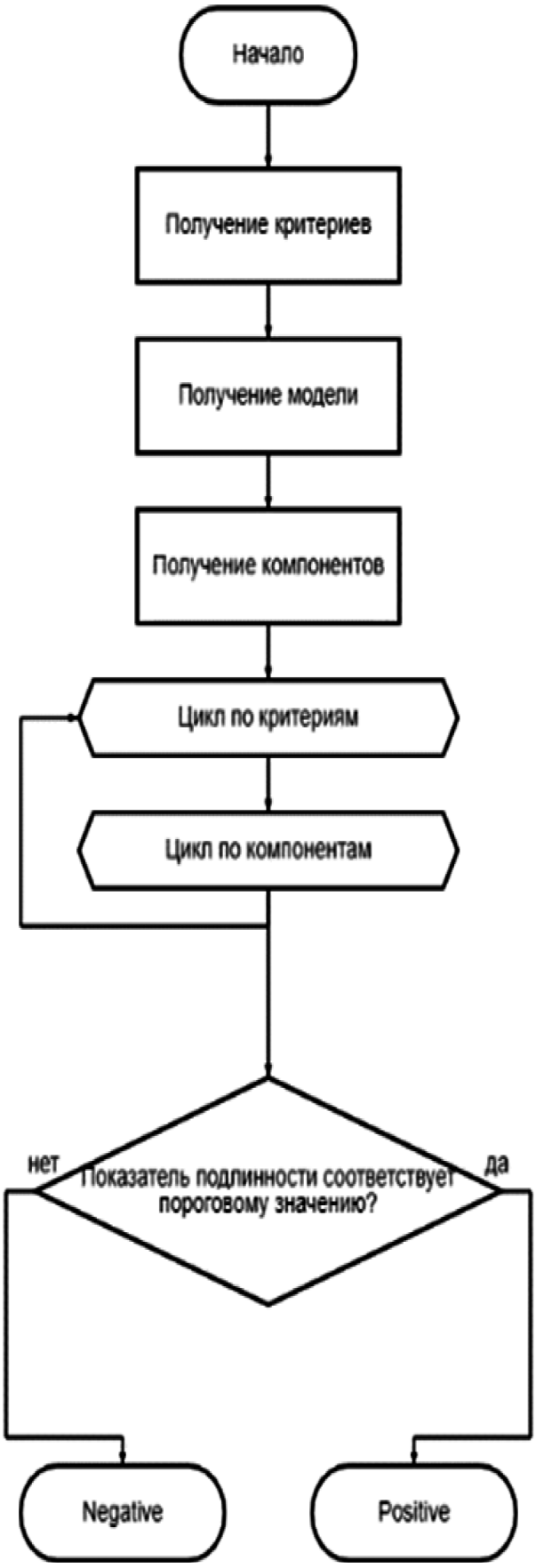

Рис. 4. Алгоритм классификации

классификации подлинности интернет — ресурсов. Было проведено исследование предметной области, обзор существующих алгоритмов машинного обучения и перечень подходов к использованию. Все методы имеют свои достоинства в рамках решения различных задач. Наиболее эффективным методом классификации интернет-ресурсов можно отметить агрегированный, комплексный подход с использованием наиболее популярных и эффективных критериев и методов.

ЛИТЕРАТУРА

1. Rao R.S., Pais A.R. Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach // Journal of Ambient Intelligence and Humanized Computing, 2019.
2. Rao R.S., Vaishnavi T., Pais A.R. CatchPhish: detection of phishing websites by inspecting URLs // Journal of Ambient Intelligence and Humanized Computing, Vol. 11, No. 2, February 2020. с. 813–825.
3. APWG. Phishing Attack Trends Report — 2Q 2020 [Электронный ресурс] [2021]. URL: https://docs.apwg.org/reports/apwg_trends_report_q4_2020.pdf (дата обращения: June.10.2021).
4. Mitchell T.M. Machine Learning. McGraw-Hill Education, 1997. 432 с.
5. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. погружение в мир нейронных сетей. Питер СПб: Питер, 2018. 480 с.
6. Журавлев Ю.И., Рязанов В.В., Сенько О.В. «Распознавание». Математические методы. программная система. Практические применения. Москва: ФАЗИС, 2006. 176 с.
7. Стрижев В.В. Методы индуктивного порождения регрессионных моделей. Москва: вычислительный центр РАН, 2008. 61 с.
8. Goodfellow I., Bengio Y., Courville A. Deep learning. The MIT Press, 2016. 800 с.
9. Шахиди А. Деревья решений: общие принципы. 2019.
10. Rocca B. Ensemble methods: bagging, boosting and stacking [Электронный ресурс] [2019]. URL: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (дата обращения: 10.июнь.2021).
11. loginom. Нормализация данных URL: https://wiki.loginom.ru/articles/data-normalization.html (дата обращения: 10.Июнь.2021).
12. Sahu B., Dehuri S., Jagadev A. A Study on the Relevance of Feature Selection Methods in Microarray Data // The Open Bioinformatics Journal, Vol. 11, 2018. с. 117–139.
13. Dimension Reduction Techniques with Python URL: https://towardsdatascience.com/dimension-reduction-techniques-with-python-f36ca7009e5c (дата обращения: 10.Июнь.2021).
14. Bergstra J., Bengio Y. Random Search for Hyper-Parameter Optimization // Journal of Machine Learning Research 13, 2012.
15. Dewancker I., McCourt M., Clark S. Bayesian Optimization for Machine Learning: A Practical Guidebook 2016. URL: https://static.sigopt.com/b/20a144d208ef255d3b981ce419667ec25d8412e2/static/pdf/SigOpt_Bayesian_Optimization_Primer.pdf (дата обращения: 10.Май.2021).
16. Дудченко П.В. Метрики оценки классификаторов в задачах медицинской диагностики // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых учёных. Томск. 2019. с. 164–165.