Анализ и формирование наборов данных сетевого трафика для обнаружения компьютерных атак

В.В. Чаругин, А.Н. Чесалин

***Аннотация* – В работе проводится анализ наборов данных сетевого трафика и формируются информативные признаки для обнаружения компьютерных атак.**

В статье рассматриваются наборы данных сетевых атак NSL-KDD и UNSW-NB15, и у них определяются избыточные признаки сетевого трафика. Осуществляется отбор наиболее значимых признаков для выявления аномалий. Формируется новый набор современных сетевых атак для тестирования алгоритмов машинного обучения.

Проводится анализ методов машинного обучения (классификатор k-ближайших соседей, классификатор случайного леса, классификатор многослойного персептрона, XGBoost) для задачи обнаружения вторжений на основе исследуемых и созданного наборов данных. Производится оценка качества классификации с использованием метрик: Accuracy и F1-score.

Результаты, полученные в данной работе, могут быть применены для тестирования, методов машинного обучения и разработки систем обнаружения вторжений.

***Ключевые слова* – сетевой трафик, наборы данных NSL-KDD, UNSW-NB15, классификатор k-ближайших соседей, классификатор случайного леса, классификатор многослойного персептрона, XGBoost, бинарная классификация, многоклассовая классификация, обнаружение вторжений, система обнаружения вторжений.**

I. Введение

Угрозы компьютерной безопасности становятся все более серьезными из-за растущих возможностей злоумышленников, влияющих на надежность передачи данных в сети. Основная цель информационной безопасности – обеспечить целостность, конфиденциальность и доступность за счет реализации различных инструментов безопасности, которые могут защитить данные и обнаружить атаки. Вторжение пытается нарушить одну из целей безопасности и заражает системы. Таким образом, появляется необходимость в создании системы обнаружения вторжений. Для построения современных систем обнаружения вторжений, способных выявлять атаки нулевого дня применяются методы искусственного интеллекта. Для работы таких систем необходимо обучить систему, для этого используют наборы сетевых данных, представляющие собой как нормальный, так и аномальный траффик.

В данной работе рассматриваются наборы данных сетевых атак NSL-KDD и UNSW-NB15. В результате анализа производится выбор значимых признаков сетевого трафика для построения нового набора данных. Выполняется тестирование алгоритмов машинного обучения на реальном сетевом трафике.

II. Исследование наборов данных сетевых атак NSL-KDD и UNSW-NB15

В качестве первого набора данных рассмотрен набор данных NSL-KDD [1, 2], построенный на основе базы KDD-99 по инициативе американской Ассоциации перспективных оборонных научных исследований DARPA, которая охватывает широкий спектр различных вторжений. В

таблице 1 представлены признаки трафика набора данных NSL-KDD с важностью в процентном соотношении.

Таблица 1 – Признаки трафика набора данных NSL-KDD с важностью в процентном соотношении.

№	Имя параметра	T	Описание	Важность параметра (%)
1	duration	I	Время продолжительности подключения	2
2	protocol_type	S	Протокол, используемый при подключении	5
3	service	S	Сетевая служба, используемая подключением	8
4	flag	S	Статус соединения – нормальное или с ошибкой	1
5	src_bytes	I	Количество отправленных байт за одно соединение	6
6	dst_bytes	I	Количество принятых байт за одно соединение	6
7	land	I	Если ip-адреса хоста источника и назначения равны, и аналогичная ситуация с портами, то параметр принимает значение 1, иначе 0	5
8	wrong_fragment	I	Общее число неверных фрагментов за это подключение	5
9	urgent	I	Количество urgent-пакетов в этом подключении	1
10	hot	I	Количество hot-индикаторов, например таких как: вход в системные директории, создание программ, выполнение программ.	2
11	num_failed_logins	I	Количество неудачных попыток входа	5
12	logged_in	B	Логин статус. 1 – если успешно вошли в систему, иначе 0	4
13	num_compromised	I	Число скомпрометированных состояний	1
14	root_shell	B	1, если root-права получены, иначе 0	5
15	su_attempted	B	1, если su root-права получены, иначе 0	1
16	num_root	I	Число root-доступов	4
17	num_file_creations	I	Число операций по созданию файлов во время соединения	1
18	num_shells	I	Число вызовов shell-оболочки	1
19	num_access_files	I	Число операций по получению контроля доступа к файлам	1
20	num_outbound_cmds	I	Число исходящих команд в FTP-сессии	1
21	is_hot_login	B	1, если логин принадлежит hot-листу т.е. если является root или администратором, иначе 0	1
22	is_guest_login	B	1, если логин является гостевым, иначе 0	1
23	count	I	Количество подключений к одному и тому же хосту назначения за последние две секунды	2
24	serror_rate	F	Процент соединений с хостом из count с SYN-ошибками	1
25	rerror_rate	F	Процент соединений с хостом из count с REJ-ошибками	1
26	same_srv_rate	F	Процент соединений с хостом из count использующих одни и те же службы	1
27	diff_srv_rate	F	Процент соединений с хостом используя разные службы	1
28	srv_count	I	Число соединений с одной и той же службой за последние две секунды.	2
29	srv_serror_rate	F	Процент соединений с SYN-ошибками при соединении по службе из srv_count	1
№	Имя параметра	T	Описание	Важность (%)
30	srv_rerror_rate	F	Процент соединений с REJ-ошибками при соединении по службе из srv_count	3
31	srv_diff_host_rate	F	Процент соединений с разными хостами при соединении по службе из srv_count	3
32	dst_host_count	I	Число соединений с тем же самым ip-адресом хоста назначения	1
33	dst_host_srv_count	I	Число соединений с тем же самым номером порта	1
34	dst_host_same_srv_rate	F	Процент соединений по той же самой службе во время соединения по ip из dst_host_count	1
35	dst_host_diff_srv_rate	F	Процент соединений по разным службам во время соединения по ip из dst_host_count	1
36	dst_host_same_src_port_rate	F	Процент соединений к тому же самому хосту приёмнику во время соединения по порту из dst_host_srv_count	3
37	dst_host_srv_diff_host_rate	F	Процент соединений с разными хостами приёмниками во время соединения по порту из dst_host_srv_count	3
38	dst_host_serror_rate	F	Процент соединений с хостом из dst_host_count с SYN-ошибками	3
39	dst_host_srv_serror_rate	F	Процент соединений с SYN-ошибками при соединении по службе из dst_host_srv_count	3
40	dst_host_rerror_rate	F	Процент соединений с хостом из dst_host_count с REJ-ошибками	1
41	dst_host_srv_rerror_rate	F	Процент соединений с REJ-ошибками при соединении по службе из dst_host_srv_count	1

В качестве второго набора данных рассмотрен набор данных UNSW-NB15 [3, 4], построенный в лаборатории Cyber Range Австралийского центра кибербезопасности.

В таблице 2 представлены признаки трафика набора данных UNSW-NB15 с важностью в процентном соотношении.

Таблица 2 – Признаки трафика набора данных NSL-KDD с важностью в процентном соотношении

№	Имя параметра	T	Описание	Важность (%)
1	id	I	Идентификатор	0
2	dur	F	Общая продолжительность	1
3	proto	S	Протокол, используемый при подключении	4
4	service	S	Сетевая служба, используемая подключением (http, ftp, ssh, …)	6
5	state	S	Состояние и зависимый от него протокол (ACC, CLO, …)	3
6	spkts	I	Количество пакетов от источника к пункту назначения	6
7	dpkts	I	Количество пакетов от пункта назначения к источнику	6
8	sbytes	I	Количество отправленных байт за одно соединение	4
9	dbytes	I	Количество принятых байт за одно соединение	4
10	rate	F	Скорость передачи и приема данных Ethernet	1
11	sttl	I	Время жизни от источника к пункту назначения	8
12	dttl	I	Время жизни от пункта назначения к источнику	8
№	Имя параметра	T	Описание	Важность (%)
13	sload	F	Количество отправленных бит в секунду	5
14	dload	F	Количество принятых бит в секунду	5

15	sloss	I	Количество отправленных пакетов, которые повторно переданы или отброшены	2
16	dloss	I	Количество принятых пакетов, которые повторно переданы или отброшены	2
17	sinpkt	F	Время прибытия между пакетами от источника (мсек)	1
18	dinpkt	F	Время прибытия между пакетами от пункта назначения (мсек)	1
19	sjit	F	Джиттер от источник (мсек)	1
20	djit	F	Джиттер от пункта назначения (мсек)	1
21	swin	I	Максимальный обьем данных, который может быть отправлен	1
22	stcpb	I	Порядковый номер источника при TCP-соединении	1
23	dtcpb	I	Порядковый номер пункта назначения при TCP соединении	1
24	dwin	I	Максимальный обьем данных, который может быть получен	1
25	tcprtt	F	Сумма ‘synack’ и ‘ackdat’ при TCP-соединении	1
26	synack	F	Время между SYN и SYN_ACK пакетами при TCP-соединении	1
27	ackdat	F	Время между SYN_ACK и SYN пакетами при TCP-соединении	1
28	smean	I	Среднее значение размера отправленного пакета	1
29	dmean	I	Среднее значение размера принятого пакета	1
30	trans_depth	I	Глубина подключения транзакций http-запроса / ответа	1
31	response_body_len	I	Размер данных, передаваемых серверной службой http	1
32	ct_srv_src	I	Количество подключений, которые содержат одну и ту же службу (4) и адрес источника из 100 подключений по данным последнего времени (31)	1
33	ct_state_ttl	I	№ для каждого состояния (5) в соответствии с конкретным диапазоном значений времени жизни источника / назначения (11) (12)	2
34	ct_dst_ltm	I	Количество подключений с одним и тем же адресом назначения в 100 подключениях по данным последнего времени (31)	1
35	ct_src_dport_ltm	I	Количество подключений с одним и тем же адресом источника и портом назначения в 100 подключениях по данным последнего времени (31)	3
36	ct_dst_sport_ltm	I	Количество подключений с одним и тем же адресом назначения и исходным портом в 100 подключениях по данным последнего времени (31)	2
37	ct_dst_src_ltm	I	Количество подключений одного и того же адреса источника и получателя в 100 подключениях по данным последнего времени (31)	1
38	is_ftp_login	B	Если доступ к сеансу ftp осуществляется пользователем и паролем, то 1, иначе 0	1
39	ct_ftp_cmd	I	Количество потоков, для которых есть команда в сеансе ftp	1
40	ct_flw_http_mthd	I	Количество потоков, у которых есть такие методы, как Get и Post, в службе http.	1
41	ct_src_ltm	I	Количество подключений с одним и тем же адресом источника в 100 подключениях по данным последнего времени (31).	2
№	Имя параметра	T	Описание	Важность (%)
42	ct_srv_dst	I	Количество подключений, которые содержат одну и ту же услугу (4) и адрес назначения из 100 подключений по данным последнего времени (31).	3
43	is_sm_ips_ports	B	Если источник равен IP-адресам назначения и номера портов одинаковы, то 1, иначе 0	
44	attack_cat	S	Название каждой категории атаки. В этом наборе данных девять категорий (Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode и Worms).	0
45	label	B	Если есть атака, то 1, иначе 0	0

III. ФОРМИРОВАНИЕ НОВОГО НАБОРА ДАННЫХ

Приведенные исследуемые наборы данных содержат избыточные атрибуты [5, 6], поэтому предлагается формирование набора данных сетевого трафика на основе выбора важных атрибутов из наборов данных NSL-KDD и UNSW-NB15, которые определяются исходя из анализа работы [7] на основе использования метода поиска ассоциативных правил.

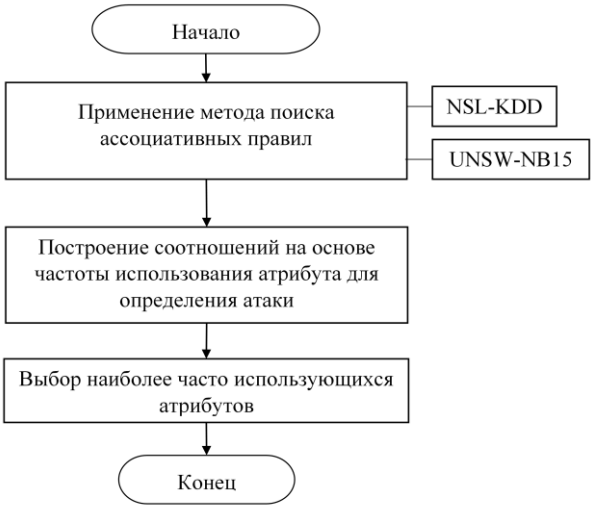

Рисунок 1 – Блок схема определения важности атрибутов набора данных

На основе анализа значений важности признаков, представленных, в таблицах 1 и 2 сформирован набор признаков, представленный в таблице 3, который используется для формирования набора данных из реального сетевого трафика.

Таблица 3 – Признаки реального сетевого трафика

№	Имя параметра	T	Описание
1	protocol_type	S	Протокол, используемый при подключении
2	service	S	Сетевая служба, используемая подключением
3	src_bytes	I	Количество отправленных байт за одно соединение
4	dst_bytes	I	Количество принятых байт за одно соединение
5	land	I	Если ip-адреса хоста источника и назначения равны, и аналогичная ситуация с портами, то параметр принимает значение 1, иначе 0
№	Имя параметра	T	Описание
6	wrong_fragment	I	Общее число неверных фрагментов за это подключение
7	num_failed_logins	I	Количество неудачных попыток входа
8	logged_in	B	Логин статус. 1 – если успешно вошли в систему, иначе 0
9	root_shell	B	1, если root-права получены, иначе 0
10	num_root	I	Число root-доступов

11	srv_rerror_rate	F	Процент соединений с REJ-ошибками при соединении по службе из srv_count
12	srv_diff_host_rate	F	Процент соединений с разными хостами при соединении по службе из srv_count
13	dst_host_same_src_port_rate	F	Процент соединений к тому же самому хосту приёмнику во время соединения по порту из dst_host_srv_count
14	dst_host_srv_diff_host_rate	F	Процент соединений с разными хостами приёмниками во время соединения по порту из dst_host_srv_count
15	dst_host_serror_rate	F	Процент соединений с хостом из dst_host_count с SYN-ошибками
16	dst_host_srv_serror_rate	F	Процент соединений с SYN-ошибками при соединении по службе из dst_host_srv_count
17	spkts	I	Количество пакетов от источника к пункту назначения
18	dpkts	I	Количество пакетов от пункта назначения к источнику
19	sttl	I	Время жизни от источника к пункту назначения
20	dttl	I	Время жизни от пункта назначения к источнику
21	sload	F	Количество отправленных бит в секунду
22	dload	F	Количество принятых бит в секунду

Для формирования нового набора данных было проведено моделирование современных сетевых атак. Смоделированный набор тренировочных данных включает 1084 записи, из них 400 нормальных, 300 SYN-Flood, 28 UDP-Flood, 300 Nmap, 36 Ipsweep, 20 Slowloris [15]. Набор тестовых данных включает 281 запись, из них 100 нормальных, 80 SYN-Flood, 7 UDP-Flood, 80 Nmap, 9 Ipsweep, 5 Slowloris. Разработанный набор данных выложен на портале Github: https://github.com/val-ugs/UDevMe.IDS/tree/Valugs/IDS.DataAccess.CSV/CsvData

IV. Тестирование методов машинного обучения на основе наборов данных

В работе проведен анализ следующих алгоритмов машинного обучения, показывающих высокое качество классификации при анализе разнородных данных:

- классификатор k-ближайших соседей; [8]
- классификатор случайного леса; [9]
- классификатор многослойного персептрона; [10]
- XGBoost. [11]

При анализе качества классификации использовались две метрики: accuracy и F1-метрика.

Метрика Accuracy — это отношение количества правильных прогнозов к общему количеству образцов. Метрика рассчитывается по формуле.

$$Accuracy = \frac{\text{число правильных прогнозов}}{\text{Общее количество образцов}} \quad (1)$$

F1 score — гармоническое среднее между точностью (Precision) и полнотой (Recall).

Precision — количество полученных от классификатора положительных ответов, являющиеся правильными.

Recall — способность классификатора предсказать как можно большее число положительных ответов из ожидаемых. [11], [12]

В таблице 4.1 представлена оценка бинарной классификации.

Таблица 4.1 – Оценка бинарной классификации.

Классы	$class_1$(Positive)	$class_2$(Negative)
$class_1$(Positive)	TP(True Positive)	FP(False Positive)
$class_2$(Negative)	FN(False Negative)	TN(True Negative)

$Precision$ рассчитывается по формуле:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$Recall$ рассчитывается по формуле:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$F1\_score$ рассчитывается по формуле:

$$F1\_score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

В таблице 4.2 представлены оценки многоклассовой классификации.

Таблица 4.2 – Оценки многоклассовой классификации

Классы	$class_1$	$class_2$	…	$class_{n-1}$	$class_n$
$class_1$	$A_{1\,1}$ (true)	$A_{1\,2}$ (error)	… (errors)	$A_{1\,n-1}$ (error)	$A_{1\,n}$ (error)
$class_2$	$A_{2\,1}$ (error)	$A_{1\,2}$ (true)	… (errors)	$A_{2\,n-1}$ (error)	$A_{2\,n}$ (error)
…	… (errors)	… (errors)	… (true при i=j, иначе error)	… (error)	… (error)
$class_{n-1}$	$A_{n-1\,n-1}$ (error)	$A_{n-1\,2}$ (error)	… (errors)	$A_{n-1\,n-1}$ (true)	$A_{n-1\,n}$ (error)
$class_n$	$A_{n\,1}$ (error)	$A_{n\,2}$ (error)	…(errors)	$A_{n\,n-1}$ (error)	$A_{n\,n}$ (true)

где A – матрица неточностей, n – количество меток (классов).

$Precision_i$ рассчитывается по формуле:

$$Precision_i = \frac{A_{i,i}}{\sum_{j=1}^{n} A_{i,j}} \quad (5)$$

где $i = 1, \dots, n$.

$Recall_i$ рассчитывается по формуле:

$$Recall_i = \frac{A_{i,i}}{\sum_{j=1}^{n} A_{j,i}} \quad (6)$$

где $i = 1, \dots, n$.

$F1\_score_i$ рассчитывается по формуле:

$$F1\_score_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (7)$$

где $i = 1, \dots, n$.

$F1\_score$ рассчитывается по формуле:

$$F1\_score = \frac{1}{n}\sum_{i=0}^{n} F1\_score_i \quad (8)$$

В таблице 5 представлены результаты классификации для набора данных NSL-KDD.

Таблица 5 – Результаты классификации для набора данных NSL-KDD

Метод	Бинарное/ Многоклассовое разделение	Метрика		Время, затраченное на обучение (час:мин:сек: мсек)	Время, затраченное на прогнозирование (час:мин:сек:мсек)
		Accuracy	F1 score		
Классификатор k-ближайших соседей	Бинарное	0.776	0.775	-	00:00:09:62
	Многоклассовое	0.711	0.242	-	00:00:10:28
Классификатор многослойного персептрона	Бинарное	0.757	0.756	00:02:01:37	00:00:00:08
	Многоклассовое	0.693	0.191	00:02:41:62	00:00:00:08
Классификатор случайного леса	Бинарное	0.786	0.784	00:13:44:45	00:00:00:01
	Многоклассовое	0.679	0.151	00:13:16:79	00:00:00:01
XGBoost	Бинарное	0.796	0.796	00:05:04:46	00:00:00:01
	Многоклассовое	0.704	0.246	00:35:15:32	00:00:00:02

В таблице 6 представлены результаты классификации набора данных UNSW-NB15.

Таблица 6 – Результаты классификации для набора данных UNSW-NB15

Метод	Бинарное/ Многоклассовое разделение	Метрика		Время, затраченное на обучение (час:мин:сек: мсек)	Время, затраченное на прогнозирование (час:мин:сек:мсек)
		Accuracy	F1 score		
Классификатор k-ближайших соседей	Бинарное	0.904	0.553	-	00:00:13:62
	Многоклассовое	0.55	0.326	-	00:00:13:27
Классификатор многослойного персептрона	Бинарное	0.91	0.574	00:01:50:21	00:00:00:06
	Многоклассовое	0.506	0.267	00:02:21:34	00:00:00:07
Классификатор случайного леса	Бинарное	0.91	0.567	00:12:46:83	00:00:00:01
	Многоклассовое	0.42	0.161	00:12:42:04	00:00:00:01
XGBoost	Бинарное	0.91	0.567	00:04:15:54	00:00:00:01
	Многоклассовое	0.605	0.346	00:22:06:88	00:00:00:01

В таблице 7 – представлены результаты классификации для набора данных сетевого трафика.

Таблица 7 – Результаты классификации на разработанном наборе данных

Метод	Бинарное/ Многоклассовое разделение	Метрика		Время, затраченное на обучение (час:мин:сек: мсек)	Время, затраченное на прогнозирование (час:мин:сек: мсек)
		Accuracy	F1 score		
Классификатор k-ближайших соседей	Бинарное	0.997	0.998	-	00:00:00:38
	Многоклассовое	0.979	0.867	-	00:00:00:42
Классификатор многослойного персептрона	Бинарное	0.925	0.922	00:00:14:57	00:00:00:01
	Многоклассовое	0.922	0.482	00:00:18:80	00:00:00:01
Классификатор случайного леса	Бинарное	0.993	0.992	00:00:05:13	00:00:00:01
	Многоклассовое	0.982	0.789	00:00:04:70	00:00:00:01
XGBoost	Бинарное	0.999	0.999	00:00:05:69	00:00:00:01
	Многоклассовое	0.999	0.999	00:00:17:75	00:00:00:01

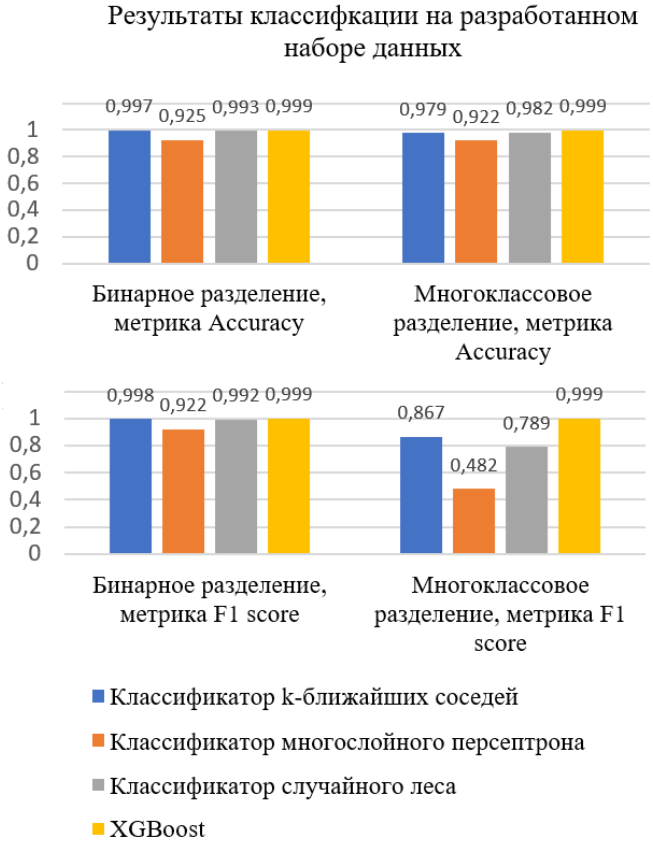

Рисунок 2 – Результаты классификации на разработанном наборе данных

Результаты анализа показывают, что в случае задачи бинарной классификации все рассматриваемые алгоритмы имеют схожее качество классификации, как по метрике accuracy, так и F1. В случае многоклассовой классификации наивысшее качество как по метрике accuracy, так и F1 получено при использовании алгоритма XGBoost.

V. Заключение

В работе рассмотрены наборы данных сетевых атак NSL-KDD и UNSW-NB15. Определены важные атрибуты наборов данных

и на основе них сформирован новый набор данных, полученный моделированием современных сетевых атак.

Проведено тестирование методов машинного обучения на наборах данных NSL-KDD, UNSW-NB15 и сформированного набора данных. Наилучшее качество классификации показал алгоритм XGBoost.

Полученные результаты могут быть применены для тестирования методов машинного обучения и разработки систем обнаружения вторжений.

Библиография

[1] L.Dhanabal, Dr. S.P. Shantharajah “A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms”, IJARCCE, vol. 4, no. 6, 2015.
[2] NSL-KDD dataset. Available: https://www.unb.ca/cic/datasets/nsl.html (URL)
[3] N. Moustafa, J. Slay. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015.
[4] The UNSW-NB15 Dataset. Available: https://research.unsw.edu.au/projects/unsw-nb15-dataset (URL)
[5] Adetunmbi Adebayo O., Adeola Oladele Stephen “ Analysis of KDD ’99 Intrusion Detection Dataset for Selection of Relevance Features”, Proceedings of the World Congress on Engineering and Computer Science, 2010.
[6] Kanimozhi V., Jacob P. UNSW-NB15 “Dataset Feature Selection and Network Intrusion Detection using Deep Learning”, International Journal of Recent Technology and Engineering, vol. 7, no. 5, 2019.
[7] Moustafa N., Slay J. “The Significant Features of the UNSW-NB15 and the KDD99 Data Sets for Network Intrusion Detection Systems”, International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, no.4, 2015.
[8] Oreshkov V. Klassifikacija dannyh metodom k-blizhajshih sosedej, 2021. Available: https://loginom.ru/blog/knn (URL)
[9] Realizacija i razbor algoritma «sluchajnyj les» na Python. Perevody, 2019. Available: https://tproger.ru/translations/python-random-forest-implementation (URL)
[10] Tarik Rashid. Sozdaem nejronnuju set'. - SPb.: OOO «Al'fa-kniga», 2017. - 272 s.
[11] Introduction to Boosted Trees. Available: //xgboost.readthedocs.io/en/stable/tutorials/model.html (URL)
[12] Sheluhin O.I. i dr. Obnaruzhenie vtorzhenij v komp'juternye seti (setevye anomalii). Uchebnoe posobie dlja vuzov – M.: Gorjachaja linija – Telekom, 2018. – 220 s: il.
[13] Chesalin A.N., Grodzenskij S.Ja., Nilov M.Ju., Agafonov A.N. Modifikacija algoritma WaldBoost dlja povyshenija jeffektivnosti reshenija zadach raspoznavanija obrazov v real'nom vremeni // Rossijskij tehnologicheskij zhurnal. 2019. T. 7. # 5. S. 20–29. Available: https://doi.org/10.32362/2500-316X-2019-7-5-20-29 (URL).
[14] Chesalin A.N. Primenenie kaskadnyh algoritmov klassifikacii dlja sovershenstvovanija sistem obnaruzhenija vtorzhenij // Nelinejnyj mir. 2022. T. 20. # 1. S. 24−41. Available: https://doi.org/10.18127/j20700970-202201-03 (URL)
[15] Samoshkin D. Perspektivnye DDoS-ataki: o chjom nuzhno znat' i kak gotovit'sja? 2020. Available: https://www.comnews.ru/content/208080/2020-07-15/2020-w29/perspektivnye-ddos-ataki-o-chyom-nuzhno-znat-i-kak-gotovitsya (URL)

Статья получена: 17 марта 2023 г.
Чаругин Валентин Валерьевич, студент РТУ МИРЭА (email: valentin.1999@mail.ru)
Александр Николаевич Чесалин к.т.н., заведующий кафедрой компьютерной и информационной безопасности, ИИИ, РТУ МИРЭА, Москва, Россия (e-mail: chesalin@mirea.ru).

Analysis and creation of network traffic datasets to detect computer attacks

V.V. Charugin, A.N. Chesalin

***Abstract* – The paper examines analysis and formation features of network traffic to detect network anomalies.**

The paper considers the NSL-KDD and UNSW-NB15 network attack datasets and identifies redundant features of network traffic in them. The selection of the most significant features is carried out to identify anomalies. A new set of modern network attacks is being formed to test machine learning algorithms.

The analysis of machine learning methods (classifier of k-nearest neighbors, classifier of random forest, classifier of multilayer perceptron, XGBoost) is carried out for the problem of intrusion detection based on the studied and created datasets. The classification quality is evaluated using the following metrics: Accuracy and F1-score.

The results obtained in this work can be applied to testing, machine learning methods and the development of intrusion detection systems.

***Keywords* – network traffic, dataset, NSL-KDD, UNSW-NB15, k-nearest neighbor classifier, random forest classifier, multilayer perceptron classifier, XGBoost, binary classification, multiclass classification, intrusion detection, intrusion detection system.**

REFERENCES

[1] L.Dhanabal, Dr. S.P. Shantharajah "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", *IJARCCE*, vol. 4, no. 6, 2015.

[2] NSL-KDD dataset. Available: https://www.unb.ca/cic/datasets/nsl.html (URL)

[3] N. Moustafa, J. Slay. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015.

[4] The UNSW-NB15 Dataset. Available: https://research.unsw.edu.au/projects/unsw-nb15-dataset (URL)

[5] Adetunmbi Adebayo O., Adeola Oladele Stephen " Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features", *Proceedings of the World Congress on Engineering and Computer Science*, 2010.

[6] Kanimozhi V., Jacob P. UNSW-NB15 "Dataset Feature Selection and Network Intrusion Detection using Deep Learning", *International Journal of Recent Technology and Engineering*, vol. 7, no. 5, 2019.

[7] Moustafa N., Slay J. "The Significant Features of the UNSW-NB15 and the KDD99 Data Sets for Network Intrusion Detection Systems", *International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, no.4, 2015.

[8] Oreshkov V. Data classification by k-nearest neighbors. 31 May 2021. Available: https://loginom.ru/blog/knn (URL) (In Russian)

[9] Implementation and analysis of the random forest algorithm in Python. Translations, 14 June 2019. Available: https://tproger.ru/translations/python-random-forest-implementation (URL) (In Russian)

[10] Tariq Rashid. Create a neural network. – St. Peterburg.: OOO «Альфа-книга», 2017. - 272 c. (In Russian)

[11] Introduction to Boosted Trees. Available: //xgboost.readthedocs.io/en/stable/tutorials/model.html (URL)

[12] Shelukhin O.I. Detection of intrusions into computer networks (network anomalies). Textbook for universities – M.: Hotline – Telecom, 2018. – 220 c: ил. (In Russian)

[13] Chesalin A.N., Grodzensky S.Ya., Nilov M.Yu., Agafonov A.N. "Modification of the WaldBoost algorithm to improve the efficiency of solving real-time pattern recognition problems", *Russian Technological Journal*, 2019, vol. 7, no 5, pp. 20–29. Available: https://doi.org/10.32362/2500-316X-2019-7-5-20-29 (URL) (In Russian)

[14] Chesalin A.N. "Application of cascade classification algorithms to improve intrusion detection systems", *Nonlinear world*, 2022, vol. 20, no. 1, pp. 24−41. Available: https://doi.org/10.18127/j20700970-202201-03 (URL). (In Russian)

[15] Samoshkin D. Promising DDoS attacks: what you need to know and how to prepare? Available: https://www.comnews.ru/content/208080/2020-07-15/2020-w29/perspektivnye-ddos-ataki-o-chyom-nuzhno-znat-i-kak-gotovitsya (URL) (In Russian)