DOI 10.37882/2223-2966.2023.8-2.02

СИСТЕМА ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ НА ОСНОВЕ ПРОЦЕССА ВЫБОРА ФУНКЦИЙ НА ОСНОВЕ АНСАМБЛЯ

AN INTRUSION DETECTION SYSTEM BASED ON AN ENSEMBLE-BASED FEATURE SELECTION PROCESS

A. Alshaibi
M. Al-Ani
A. Al-Azzawi
A. Konev

Summary. The article analyzes software that provides security monitoring and real-time threat detection. It is shown that most of the methods have a large intrusion detection time due to the high traffic noise. A new method for selecting intrusion signs based on obtaining information, its classification and creating options using the particle swarm method is proposed. This method consists in a preliminary selection of sample groups characteristic of various invasion methods. Based on this method, a technological scheme for intrusion detection and a platform for pre-configuring an intrusion detection detector for the main groups of intrusions and attacks are proposed. Experimental data comparing the proposed method with the reference method without assembly showed its efficiency almost two times higher. In conclusion, it is noted that the possibility of using this method in the field of cloud computing. The high classification accuracy of ensemble classifiers compared to a single machine learning algorithm for detecting IDS in the cloud shows great promise for this method.

Keywords: information protection, security, monitoring, intrusion detection, network anomalies, anomaly assembly.

Алшаиби Ахмед Джамал
аспирант, Томский государственный университет систем управления и радиоэлектроники
ahmed.jamal.alshaibi88@gmail.com
Аль-Ани Мустафа Мажид
аспирант, Томский государственный университет систем управления и радиоэлектроники
safo.alany@gmail.com
Аль-Азави Абир Ясин
аспирант, Томский государственный университет систем управления и радиоэлектроники
abeerda89@gmail.com
Конев Антон Александрович
к.т.н., доцент, Томский государственный университет систем управления и радиоэлектроники
kaa@fb.tusur.ru

Аннотация. В статье анализируется программное обеспечение, которое предоставляет возможности мониторинга безопасности, и обнаружения угроз в режиме реального времени. Показано, что большинство методов обладают большим временем обнаружения вторжения из-за большой зашумленности трафика. Предложен новый метод выбора признаков вторжения на основе получения информации, ее классификации и создания опций по методу роя частиц. Этот метод заключается в предварительной подборке эталонов групп, характерных для различных методов вторжения. На основе этого метода предложена технологическая схема обнаружения вторжения и платформа для предварительной настройки детектора обнаружения вторжения на основные группы вторжений и атак. Экспериментальные данные сравнения предложенного метода с эталонным методом без ассемблирования показали его эффективность выше почти в два раза. В заключение отмечается, возможность применения данного метода в сфере облачных вычислений. Высокая точность классификации ансамблевых классификаторов по сравнению с единым алгоритмом машинного обучения для обнаружения IDS в облаке показывает большие перспективы этого метода.

Ключевые слова: защита информации, безопасность, мониторинг, обнаружение вторжения, сетевые аномалии, ассемблирование аномалий.

Введение

Одна из основных проблем исследования безопасности, с которыми сталкиваются традиционные методы системы обнаружения вторжения (Intrusion Detection System — IDS) заключается в их неспособности обрабатывать большие объемы сетевых данных и обнаруживать современные кибератаки с высокой точностью обнаружения и низким уровнем ложных срабатываний [1–3]. Следовательно, существует потребность в эффективных и надежных схемах IDS, которые могут справиться с этими постоянно меняющимися парадигмами кибербезопасности [4, 5, 6]. Таким образом, методы машинного обучения становятся очень популярными в разработке современных систем обнаружения вторжений [7, 8, 10].

Алгоритмы машинного обучения широко используются для вторжений обнаружения [11–14]. Однако исследования доказали, что производительность нескольких классификаторов на основе IDS намного лучше, чем один классификатор IDS, что побудило разработать ансамблевую модель обнаружения вторжений.

Материалы и методы

В статье предлагается новый метод на основе выбора подбора функций, которые в дальнейшем могут быть ис-

пользованы для обнаружения вторжения. Тестирование набора производится на основе выбора эталонных наборов данных CICIDS-2017, предоставленном Университетом Нью-Брансуика [9].

Выбор функций с использованием методов машинного обучения

Выбор признаков — важный шаг, на котором используются классификаторы машинного обучения для процесса обнаружения вторжений и распознавания образов.

Алгоритмы выбора признаков классифицируются на три группы, а именно фильтр, оболочка и гибрид (рис. 1).

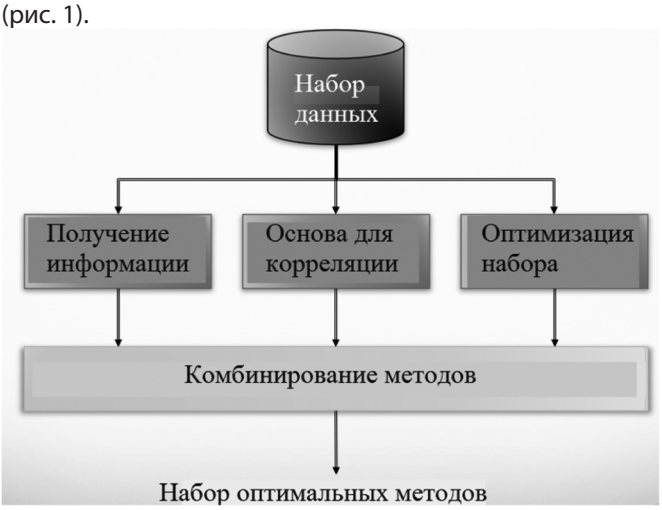

Рис. 1. Предлагаемый подход к выбору признаков на основе ансамбля

Предлагаемый метод выбора признаков использует три метода выбора признаков:

— выбора признаков на основе получения информации;
— характеристика на основе корреляции;
— метод оптимизации роя частиц для создания лучшего подмножества функций.

Каждый из этих методов генерирует подмножество наилучших возможных функций, используя образец набора данных, который был создан после фильтрации данных и нормализации данных.

Платформа обнаружения вторжений для обнаружения сетевых аномалий на основе ансамбля

В этом подразделе представим платформу, основанную на ансамбле схем обнаружения вторжения для сети (Рис 2).

Модуль выбора признаков на основе ансамбля использует три метода машинного обучения для создания отдельных подмножеств функций. Эти три отдельных подмножества функций дополнительно объединяются с использованием метода объединения подмножеств и оптимальное подмножество признаков на основе ансамбля для набора данных.

Был подготовлен набор данных, состоящий из восьми файлов, всего 2830743 записи. Каждая запись описывается 78 признаками. Это помеченный набор данных при этом каждая запись помечена как доброкачественная или как одна из атак из четырнадцати категорий угроз.

Поскольку набор данных был создан путем захвата сетевых данных из различных источников, он состоит из зашумленных и избыточных данных, которые нужно фильтровать вначале во избежание ошибок на точность обнаружения. Для этого была произведена фильтрация данных. В результате были подобраны наборы, которые являются характерными для определенных видов атак (таблица 1).

Таблица 1.

Наборы записей после фильтрации

Класс набора записей	Общее количество записей
Доброкачественное	188 955
Бот	1956
DDoS-атаки	99 999
Портскан	158 800
Веб-атака BruteForce	1 507
Веб-атака — XSS	652
Веб-атака — SQL-инъекция	21
FTP — Пататор	7 935
SSH — Пататор	5 897
DoS Slowloris	5796
DoS Slowhttptest	5 499
DoS Goldeneye	10 293
Всего записей	576 061

По наборам, данным в результате подбора, основанного на основании нейро-сетевых алгоритмов были подобраны наборы, характерные для данного вида атак., которые представлены в таблице 2. Эти выбранные функции, основанные на ансамбле, обеспечивают представления набора данных CICIDS-2017, и они будут использованы для предлагаемого ансамбля схемы обнаружения вторжений.

В таблице 3 показано сравнение между тремя отдельными методами и особенности ансамбля.

Можно заметить, что хотя производительность модели с 27 функций IG и 17 функций ансамбля одинаковы, важно отметить здесь, что количество функций для

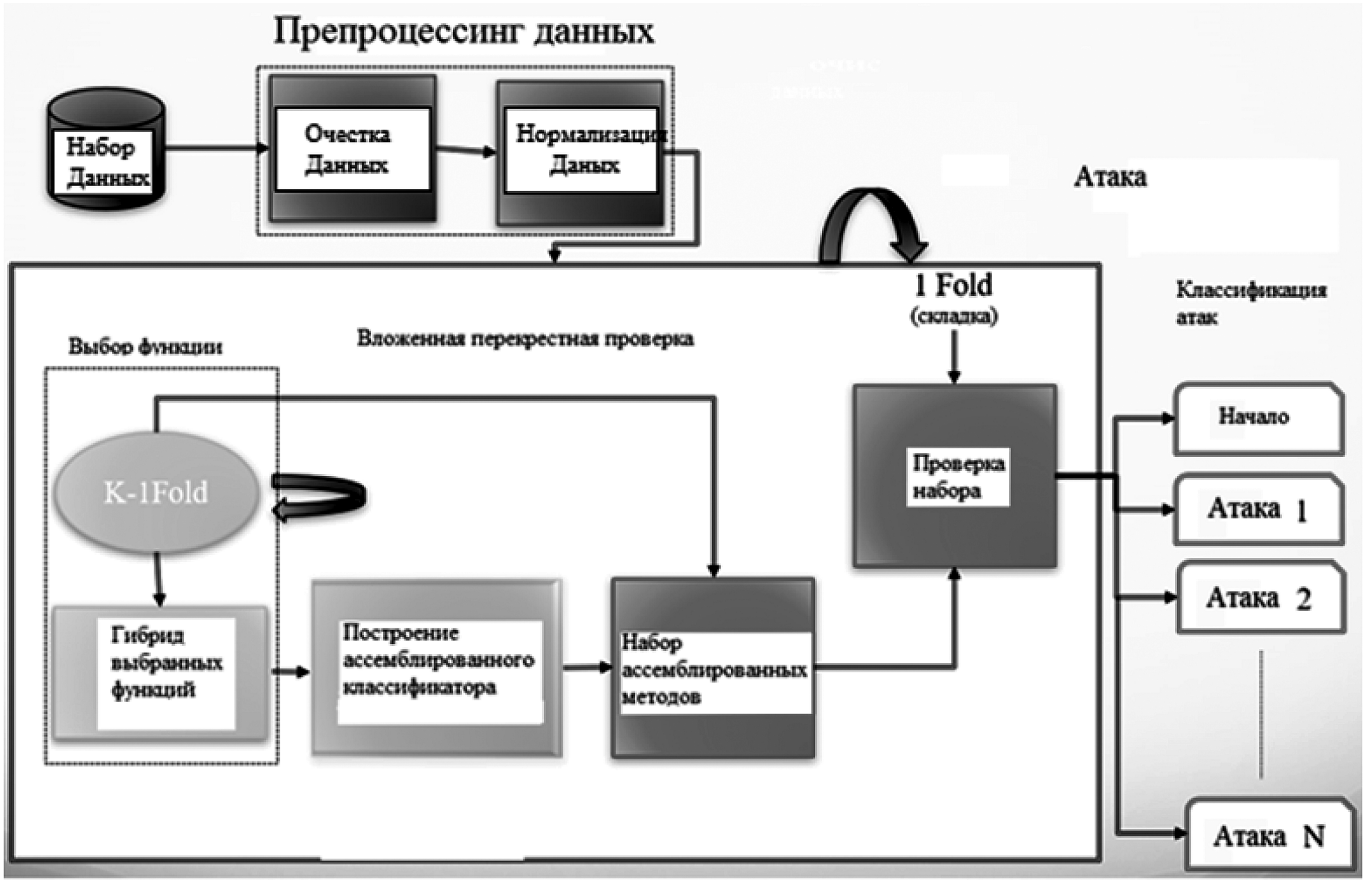

Рис. 2. Предлагаемый выбор функций ансамбля и платформа обнаружения сетевых вторжений на основе ансамбля

IG намного больше, чем функции на основе ансамбля. Следовательно, мы достигли той же точности, используя функцию на основе ансамбля.

Таблица 2.

Наборы записей после фильтрации

Номер набора функции	Название набора функции
1	Назначение — порт
7	Макс. длина пакета вперед
11	Макс. длина пакета назад
12	Минимальная длина пакета
13	Средняя длина пакета
18	Поток IAT Std
19	Расход IAT Макс
24	Передний IAT Макс
37	Максимальная длина пакета
38	Средняя длина пакета
39	Стандартная длина пакета
40	Разница в длине пакета
49	Средний размер пакета
51	Средний размер сегмента BWD
56	Init_Win_bytes_forward
57	Init_Win_bytes_backward
59	Min_seg_size_forward

Таблица 3.

Сравнительные результаты методов Feature Selection для CICIDS — 2017

FS метод	Acc	Точность	Вызов	F-среднее	MCC
IG — 27 функций	0.997	0.997	0.997	0.997	0.997
CFS — 21 функций	0.982	0.982	0.982	0.981	0.987
PSO — 13 функций	0.996	0.996	0.996	0.996	0.996
Ансамбль из 17 функций	0.997	0.997	0.997	0.997	0.997

Таким образом, мы утверждаем, наш подход FS, основанный на ансамбле, работает лучше, чем автономная индивидуальная функция схемы отбора.

В Таблице 4 и Таблице 5 представлены общие характеристики модели ансамбля для набор данных CICIDS-2017 с исходным набором признаков из 68 и уменьшенным набором функций из 17 функций, сгенерированных с использованием предложенного нами ансамблевого метода FS. Мы наблюдаем из чисел в обеих этих таблицах, точность для нашей модели ансамбля такая же, что и для трех базовых классификаторов. Таким образом, предлагаемый ансамбль IDS на основе схемы с методом выбора признаков на основе ансамбля (17 признаков), имеет высокие результаты производительности на наборе дан-

ных CICIDS — 2017. Эти результаты так же хороши, как и при использовании полного набора функций (68 функций) в предложенной схеме. Это наблюдение приводит нас к выводу, что использование нашего ансамблевого алгоритма выбора признаков, уменьшает многомерность базы данных CICIDS — 2017 без ущерба для параметров производительности.

Однако отметим, что, когда KNN используется в качестве единого классификатора для модели IDS, точность значения F-Measure и MCC довольно низкие по сравнению с предложенной нами схемы IDS на основе ансамбля.

Таблица 4.

Общая производительность модели ансамбля с набором данных CICIDS2017 (68 функций)

Классификатор	Acc	Точность	Вызов	F-среднее	MCC
C4.5	0.997	0.997	0.997	0.997	0.996
Random Forest	0.997	0.987	0.987	0.997	0.987
KNN	0.997	0.829	0.994	0.878	0.884
Ensemble	0.997	0.997	0.997	0.997	0.997

Таблица 5.

Общая производительность модели ансамбля с набором данных CICIDS2017 (17 функций)

Классификатор	Acc	Точность	Вызов	F-среднее	MCC
C4.5	0.997	0.997	0.997	0.997	0.996
Random Forest	0.997	0.987	0.987	0.997	0.987
KNN	0.995	0.831	0.995	0.827	0.836
Ensemble	0.997	0.997	0.997	0.997	0.997

Предлагаемая модель IDS на основе ансамбля реализована с использованием набора данных. Можно с уверенностью утверждать, что точность обнаружения IDS для предлагаемого ансамбля обеспечивает сравнимую точность обнаружения с меньшим, лучшим качеством подмножества признаков информативного ансамбля.

Время, затрачиваемое моделью IDS без выбора признаков, намного больше, из-за зашумленности данных. Однако модель IDS занимает почти половину времени, когда осуществляется выбор признаков. Это одно из нескольких преимуществ выбора признаков, поскольку это улучшает время, необходимое для обнаружения вторжений.

Без выбора признаков обнаружение осуществляется 839,63 секунды, с выбором признаков ансамбля — 466,78 секунды.

Выводы

В работе были выявлены три основные проблемы в области обнаружения вторжений:

— отсутствие использования ансамблевых методов для решения задачи выбора признаков, связанной с высокой размерностью наборов данных IDS.
— алгоритмы машинного обучения с одним классификатором, которые приводят к слабому классификатору и следовательно, отрицательно влияют на производительность модели IDS.
— чрезмерное использование традиционных эталонных наборов данных IDS в отсутствие современных общедоступных наборов данных проверки IDS, которые могут привести к вводящей в заблуждение оценке производительности IDS, поскольку старые наборы данных не содержат недавние атаки или современную топологию сети.

Эти пробелы в существующих исследованиях IDS были устранены следующим образом.

Были предложена, разработана и внедрена схема наборов базовых признаков ансамбля наборов данных CICIDS-2017 [9], которая используется для объединения трех методов выбора отдельных признаков: получение информации, выбор функций корреляции и оптимизация наборов методом роя частиц.

В результате проделанных экспериментальных исследований было показано, что, используя предложенную схему ансамбля 17 функций, сгенерированных нашим ансамблевым методом, дают такую же высокую точность, как и у каждого из отдельных методов, использующих более высокое количество признаков по сравнению с нашей схемой. Таким образом, модель IDS, созданная с использованием функции ансамбля имеют меньшее вычислительное время по сравнению с одиночной схемой. Кроме того, показано, что важность методов выбора признаков на основе ансамбля для уменьшения размерности усиливается на основе экспериментального результата.

При проведении исследований применялся последний доступный набор данных, который представляет сетевая настройка текущего времени, а также охватывает новейшую сложную сеть угроз, что делает его идеальным набором данных для оценки производительности IDS.

Обнаружение вторжений не только ограничивается традиционными сетями, но также играет значительную роль в современных корпоративных сетях, таких как сети облачных вычислений. Совместная IDS с корреляцией предупреждений предлагает лучшую облачную среду безопасности, чем автономные IDS. Кроме того, мы также предложили платформу обнаружения вторжения для облака. Предполагается развитие ассемблированного подхода к обнаружению в облачной архитектуре, как расширение нашего текущего предложенного метода

на будущее. Высокая точность классификации ансамблевых классификаторов по сравнению с единым алгоритмом машинного обучения для обнаружения IDS в облаке показывает большие перспективы этого метода.

Подход, представленный в данной работе, является надежным и использует методы на основе фильтров, а именно получение информации, выбор корреляционных признаков и оптимизацию роя частиц. Мы считаем, что в будущем есть много возможностей для изучения схемы на основе ансамбля для набора данных, направлены на создание подходов, которые являются надежными и генерируют ранжирование функций для лучших подмножеств функций выбранных наборов данных IDS. Ансамбльный подход, использующий комбинацию методов на основе фильтров и встроенных методов выбора признаков могут стать еще одной областью для дальнейшего изучения в будущем, чтобы построить оптимизированные подмножества функций для набора данных.

Кроме того, при ансамблевой схеме обнаружения вторжений выбор базы классификаторов очень важны.

ЛИТЕРАТУРА

1. ALJAWARNEH S., YASSEIN M. B., AND ALJUNDI M., An enhanced j48 classification algorithm for the anomaly intrusion detection systems, Cluster Computing, 22, (2019), pp. 10549–10565.
2. ALLOGHANI M., AL-JUMEILY D., MUSTAFINA J., HUSSAIN A., AND ALJAAF A. J. A systematic review on supervised and unsupervised machine learning algorithms for data science, Supervised and Unsupervised Learning for Data Science, (2020), pp. 3–21.
3. Alshaibi A.J., Al-Ani M.M., Kadum J. The effect of integration and effectiveness of artificial neural networks on information security tasks //AIP Conference Proceedings 2591 (1)
4. Alshaibi, Ahmed, Mustafa Al-Ani, Abeer Al-Azzawi, Anton Konev, and Alexander Shelupanov. 2022. "The Comparison of Cybersecurity Datasets" Data 7, no. 2: 22.
5. CUI Z., CHANG Y., ZHANG J., CAI X., AND ZHANG W. Improved NSGA-III with selection-and-elimination operator, Swarm and Evolutionary Computation, 49 (2019), pp. 23–33.
6. Jamal A.A., Majid A.AM., Konev A., Kosachenko T., Shelupanov A. A review on security analysis of cyber physical systems using Machine learning // Materials Today: Proceedings 80, 2302-2306
7. JIANG M AND XU X., Application and performance analysis of data preprocessing for intrusion detection system, in International Conference on Science of Cyber Security, Springer, 2019, pp. 163–177.
8. Majid A.AM., Alshaibi A.J., Kostyuchenko E., Shelupanov A. A review of artificial intelligence based malware detection using deep learning // Materials Today: Proceedings 80, 2678–2683
9. Nagar, U, Nanda, P & He, X Feature Analysis and Ensemble-based Intrusion Detection Scheme using CICIDS — 2017 dataset. (Submitted to Wiley, Concurrency and Computation Practice and Experience Journal.)
10. Байгутлина И.А., Замятин П.А. Геоинформационные технологии, киберспорт и кибербезопасность // Славянский форум. 2021. № 2 (32). С. 316–326.
11. Будаковский Д.В., Козин Е.А., Петраки А.В., Кондратеня К.А. Методы автоматизации поиска утечек информации в сети интернет // Наука и бизнес: пути развития. 2020. № 6 (108). С. 108–110.
12. Булгаков С.В., Ковальчук А.К., Цветков В.Я., Шайтура С.В. Защита информации в ГИС — М.: МГТУ им. Баумана, 2007.
13. Голкина Г.Е., Шайтура С.В. Безопасность бухгалтерских информационных систем — Учебное пособие — Бургас, 2016
14. Кокорева Л.А., Шайтура С.В. Безопасность платежных систем России // Славянский форум. — 2015.— № 1 (7) — с. 92–100.