УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2021 Том 21 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2021 Vol. 21 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2021-21-2-234-240
УДК 004.056

Применение бэггинга при поиске аномалий сетевого трафика

Бабыр Темирбекулы Рзаев[1]✉, Илья Сергеевич Лебедев[2]

[1] Казахский агротехнический университет им. С. Сейфуллина, Нур-Султан, 010000, Республика Казахстан
[2] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] pathinchaos@gmail.com✉, https://orcid.org/0000-0002-9671-650X
[2] isl_box@mail.ru, https://orcid.org/0000-0001-6753-2181

Аннотация

Предмет исследования. Рассмотрены подходы к решению задачи выявления аномальных ситуаций в информационно-телекоммуникационных системах, на основе методов искусственного интеллекта, анализирующих статистическую информацию пакетов трафика в различных режимах и состояниях. **Метод.** Представленный метод выявления аномальной ситуации основан на обработке полученных кортежей значений пакетов сетевого трафика путем применения бэггинга алгоритмов классификации машинного обучения. Сетевой трафик рассматривается как множество кортежей параметров пакетов, распределенных по дискретам времени. В отличие от существующих, метод не требует специальной подготовки данных, получаемые ошибки при классификации кортежей значений пакетов отдельными алгоритмами классификации усредняются их «коллективным» голосованием. Предложенное решение с целью повышения показателя точности дает возможность использовать оптимизированные для разных видов событий и аномалий алгоритмы классификации, обученные на различных обучающих выборках, представленных в виде кортежей параметров сетевых пакетов. Различность алгоритмов достигается внесением дисбаланса в обучающие выборки. **Основные результаты.** Приведено описание эксперимента с использованием алгоритмов классификации машинного обучения Naïve Bayes, Hoeffding Tree, J48, Random Forest, Random Tree и REP Tree. Оценка выполнена на открытом датасете NSL-KDD при поиске паразитного трафика. Получены результаты оценки для каждого классификатора по отдельности и с применением бэггинга классифицирующих алгоритмов. **Практическая значимость.** Метод может быть применен в системах мониторинга информационной безопасности при анализе сетевого трафика. Особенностью предлагаемого решения является возможность его масштабирования и комбинирования путем добавления новых алгоритмов классификации машинного обучения. В дальнейшем в ходе функционирования возможно вносить изменения в состав алгоритмов классификации, что позволит улучшить показатели точности идентификации потенциального деструктивного воздействия.

Ключевые слова

бэггинг, выявление аномалий, паразитный трафик, информационная безопасность

Ссылка для цитирования: Рзаев Б.Т., Лебедев И.С. Применение бэггинга при поиске аномалий сетевого трафика // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 2. С. 234–240. doi: 10.17586/2226-1494-2021-21-2-234-240

Applying bagging in finding network traffic anomalies

Babyr T. Rzayev[1]✉, Ilya S. Lebedev[2]

[1] Saken Seifullin Kazakh Agrotechnical University, Nur-Sultan, 010000, Kazakhstan
[2] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

[1] pathinchaos@gmail.com✉, https://orcid.org/0000-0002-9671-650X
[2] isl_box@mail.ru, https://orcid.org/0000-0001-6753-2181

Abstract

The authors consider approaches to solving the problem of identifying anomalous situations in information and telecommunication systems, based on artificial intelligence methods that analyze the statistical information on traffic packets in various modes and states. We propose a method for detecting an anomalous situation based on the obtained tuples of values of network traffic packets by applying bagging classifying algorithms of machine learning. The network traffic is treated as a set of tuples of packet parameters, distributed over sample time. In contrast to the existing ones, the method does not require special data preparation; the errors in the classification of tuples of package values by individual classification algorithms are averaged by "collective" voting of the classifying algorithms. The given solution to the increase of the accuracy index makes it possible to use the classifying algorithms optimized for different types of events and anomalies, trained on various training samples in the form of tuples of network packet parameters. The difference between the algorithms is achieved by introducing an imbalance to the training sets. We describe an experiment conducted by using Naïve Bayes, Hoeffding Tree, J48, Random Forest, Random Tree and REP Tree classification algorithms of machine learning. The evaluation was performed on the open NSL-KDD dataset while searching for parasitic traffic. The paper presents the results of evaluation for each classifier individually and with bagging classifying algorithms. The method can be used in information security monitoring systems to analyze network traffic. The peculiarity of the proposed solution is the possibility of scaling and combining it by adding new classifying algorithms of machine learning. In the future, in the course of operation, it is possible to make changes in the composition of the classifying algorithms, which will improve the accuracy of the identification of potential destructive impact.

Keywords

bagging, anomaly detection, parasitic traffic, information security

For citation: Rzayev B.T., Lebedev I.S. Applying bagging in finding network traffic anomalies. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no 2, pp. 234–240 (in Russian). doi: 10.17586/2226-1494-2021-21-2-234-240

Введение

Функционирование информационно-телекоммуникационных систем (ИТКС) требует постоянного мониторинга появления всевозможных сбоев, отказов, коллизий, связанных с обработкой сетевого трафика. Развитие концепции промышленного интернета, интернета вещей вызывает необходимость оценки работоспособности, функциональной безопасности отдельных сетевых устройств и образуемых ими сегментов сети. Анализ состояния осуществляется с помощью различных мониторов, обрабатывающих внутреннюю и внешнюю информацию, содержащую статистические данные сетевых пакетов и показателей их обработки. В результате появляются многомерные временные ряды и кортежи значений, которые содержат множество изменяемых во времени параметров, отражающих функционирование системы.

Рост, постоянное накопление объемов информации сетевого трафика в различных сегментах промышленных сетей и сетей интернета вещей, обуславливает необходимость применения методов искусственного интеллекта для обработки временных рядов и кортежей значений, с целью обнаружения аномалий.

В зависимости от задач анализа состояния информационной безопасности каналов связи и самой ИТКС, можно выделить ряд направлений, основанных на классификации, кластеризации и прогнозировании [1–4].

Реализация таких подходов вызывает определенные трудности. В ходе функционирования устройств интернета вещей могут возникать коллизии как на уровне информационной системы, так и отдельного сетевого сегмента или устройства, которые оказывают влияние на процессы приема и передачи информации, увеличивают загрузку каналов связи, уменьшают производительность и скорость обработки команд и сообщений [5–7]. Обнаружение и предотвращение подобных ситуаций требует совершенствования моделей, методов мониторинга состояния, позволяющих осуществлять анализ причинно-следственных связей и переходов. Адаптация методов статистического анализа, формирование прецедентных, событийных моделей дает возможность прогнозирования для предотвращения негативных последствий [5–10].

Выявление аномальной ситуации сетевого сегмента базируется на статистической информации трафика в различных режимах и состояниях и осуществляется с помощью нейронных сетей, марковских моделей, методов машинного обучения и других. Формируемые кортежи признаков имеют вид паттернов, предназначенных для анализа поведения устройства в различных режимах работы [11–14].

Основным недостатком существующих подходов является то, что системы, реализующие их, не всегда могут быть эффективны в условиях постоянных изменений конфигурации и архитектуры. В связи с этим возникает необходимость разработки и адаптации методов анализа информации, устойчивых к меняющимся условиям функционирования, обеспечивающих заданную полноту и точность выявления аномальных ситуаций.

Цель данной работы — повышение показателей качества идентификации событий в сетевом трафике за счет применения ансамбля алгоритмов, обученных на различных несбалансированных обучающих выборках. Предлагаемое решение направлено на достижение разнообразия классификаторов. Возникающий эффект разброса ответов алгоритмов сглаживается применением процедуры голосования.

Предлагаемый подход

Современный анализ сетевого трафика в большинстве случаев основывается на методах машинного обучения. Автоматическое, без участия эксперта, извлечение признаков и универсальность является несомненным положительным качеством таких подходов.

Однако, ввиду специфики области информационной безопасности, применяемые злоумышленником атаки на устройства, сети и телекоммуникации каждый раз обладают определенной «новизной» и неповторимостью. С другой стороны, увеличивается доля защищенных соединений (P2P-сервисов, HTTPS), где происходит сжатие и шифрование трафика. Следовательно, могут возникать проблемные ситуации, связанные с функционированием алгоритмов, входными параметрами, извлечением анализируемых признаков и их интерпретацией для разного рода информационных воздействий и атак. В связи с этим предлагается метод выявления аномальной ситуации информационно-телекоммуникационной системы, основанный на бэггинге, и в идеале дает возможность анализировать деструктивные воздействия на ИТКС, сопоставляя результаты нескольких классификаторов.

Формализованное описание построения ансамбля классифицирующих алгоритмов представлено в работах [11, 15–17].

В этих работах рассмотрено конечное множество l состояний системы $\{z_1, \ldots, z_l\} \in Z$, которые меняются в дискретные моменты времени под внутренними и внешними воздействиями.

Имеется выборка, где для различных состояний z_i, $i = 1, \ldots, l$ получены значения исследуемых параметров, что позволяет поставить в соответствие набор кортежей из множества X каждому состоянию.

$X_i = (x_{1i}, \ldots, x_{ni})$ содержит кортеж значений длины $n \geq 2$.

Множество состояний Z определяется кортежами $\{X_1, X_2 \ldots, X_m\} \in X$, где m — количество записей в выборке, отражающих поведение процесса в различных состояниях.

Бинарное множество классов C изначально разбитое на подмножества опасных C_1 и безопасных C_2 считается соответствующим множеству состояний.

Таким образом, имеется размеченная конечная обучающая выборка, состоящая из кортежей

$$X = \{(x_{11}, \ldots, x_{n1}), (x_{12}, \ldots, X_{n2}), \ldots, (x_{1m}, \ldots, X_{nm})\}. \quad (1)$$

Необходимо для входного кортежа значений X построить алгоритм классификации $a = \{a_1, a_2 \ldots, a_k\}$, отображающий $Z \to C$, где k — количество базовых классифицирующих алгоритмов в ансамбле.

Соответствие текущего наблюдения одному из подмножеств C_1 или C_2 определяется на основе решающего правила $\varphi_j(X_i)$ алгоритма a_j, которое вычисляется функцией $f_j(X_i)$, приводящей к разбиению пространства на две непересекающиеся области:

$$\varphi_j(X_i) = \begin{cases} C_1, \text{ при } f_j(X_i) \geq \varepsilon \\ C_2, \text{ при } f_j(X_i) < \varepsilon \end{cases}, \quad (2)$$

где ε — пороговое значение.

В задачах идентификации аномалий в трафике, вследствие особенности реализации функции $f_j(X_i)$, разделяющей пространство, для алгоритма классификации возникает ошибка, которую возможно сгладить последовательностью k независимо друг от друга обученных базовых классификаторов a_q, $q = 1, \ldots, k$.

Тогда, в отличие от [17], с учетом бинарной классификации, ответ алгоритма $a_q(X_i)$, полученный в результате работы, позволяет определить класс подмножеств $C_j, j = 1,2$, принадлежащий бинарному множеству классов опасных и безопасных состояний C.

$\{P_q(C_j|X_i)\}^2_{j=1}$ — апостериорная вероятность для q-го классификатора после обучения на выборке.

Если использовать классификаторы a_q, $q = 1, \ldots, k$ по отдельности, то вероятность принадлежности к классу поступающего на вход кортежа будет определяться выражением:

$$a_q(X_i) = \max_{j=1,2} P_q(C_j|X_i). \quad (3)$$

Результирующий класс, предсказываемый ансамблем классификаторов для кортежа X_i возможно определить на основе значений функций F_1 и F_2 для бинарного подмножества C_1 и C_2:

$$\begin{gathered} F_1(a_1(X_i), \ldots, a_k(X_i)) = \frac{1}{k}\sum_{q=1}^{k} P_q(C_1|X_i), \\ F_2(a_1(X_i), \ldots, a_k(X_i)) = \frac{1}{k}\sum_{q=1}^{k} P_q(C_2|X_i) = \\ = 1 - F_1(a_1(X_i), \ldots, a_k(X_i)). \end{gathered} \quad (4)$$

Реализация ансамбля базовых алгоритмов описывается выражением:

$$a(X_i) = \varphi(F_1(a_1(X_i), \ldots, a_k(X_i)), F_2(a_1(X_i), \ldots, a_k(X_i)) \quad (5)$$

где φ — решающее правило, позволяющее определить вероятностную оценку и установить номер класса.

Таким образом, приводимое решение использует бинарную классификацию. Формализация обобщена решающим правилом, переводящим оценку в номер класса. Для подсчета результатов рассматриваемого ансамбля используется вспомогательное множество (пространство оценок).

В целях достижения различности алгоритмов, входящих в модель, их обучение происходит независимо друга от друга как на случайно выбранных, так и на несбалансированных подмножествах обучающей выборки. В работе рассматриваются качественные показатели ансамбля, когда применяется несбалансированная выборка.

Экспериментальное исследование ансамбля классификаторов

Экспериментальная оценка рассматриваемого подхода проведена на датасете NSL-KDD [18, 19]. В рамках эксперимента осуществлена бинарная классификация состояний телекоммуникационной системы (идентификации паразитного и нормального трафика).

Одним из проблемных вопросов бэггинга стал состав ансамбля классификаторов. Существуют лишь отдельные рекомендуемые правила, связанные с формированием обучающих выборок. Большая часть исследований посвящена использованию в ансамбле «слабых» классифицирующих алгоритмов. В ряде работ в целях обеспечения условий, направленных на достижение «различности» классификаторов, рекомендовано

не использовать устойчивые к изменению обучающие подвыборки [20, 21]. Рассмотрено применение точных классификаторов совместно с относительно слабыми [22], исследованы комбинации ансамблей, в которых есть неравнозначные по качественным показателям алгоритмы [22–25]. Приведены подходы, использующие стекинг, блендинг, многоуровневый стекинг [18, 21, 22].

На сегодняшний день не существует регламентированных правил формирования состава ансамбля, что требует дополнительных исследований и экспериментов. Выбор модели происходит исходя из задачи, где состав алгоритмов определяется заранее, а его изменение будет затруднено. В случае функционирования в автономном режиме, в зависимости от условий, возникающих аномалий, возможностей по обучению и формированию обучающей выборки, классификаторы могут быть как сильными, так и слабыми. В связи с этим необходимо исследовать характеристики в различных ситуациях.

Формирование обучающей выборки — важная составляющая для разработки успешной модели [18, 26, 27]. Существуют определенные рекомендации по формированию обучающих множеств, однако на практике их выполнение вызывает затруднение. Не всегда удается получить вероятность появления объектов и признаков, равную вероятности их появления в генеральной совокупности. Проблемные вопросы формирования обучающих выборок связаны с обнаружением, разделением и правильной интерпретацией фоновых и значимых закономерностей, отсутствием обучающих объектов определенного вида и элементов признаковой системы, не точными диапазонами значений переменных, разбалансировкой, появлением внешних закономерностей, связанных с условиями формирования обучающего множества [26].

В эксперименте рассмотрен случай голосования, когда в ансамбле участвовали как сильные, так и слабые классификаторы, разнообразие которых формировалось обучающей выборкой. Оценка производилась для классификаторов: Naïve Bayes (NB), Hoeffding Tree (HT), J48, Random Forest (RF), Random Tree (RT) и REP Tree (REP). Результаты анализа получены с использованием свободно распространяемого программного обеспечения Weka.

Выборка была размечена и разделена на две части, одна из которых применялась в качестве обучающей, а другая использовалась для тестирования. Вектор структуры данных включал более 40 значений атрибутов, описание которых представлено в [19].

Полученный из датасета сетевой трафик рассмотрен как последовательность кортежей значений (1) разнородных пакетов, около 50 % из которых отражает обычный трафик и 50 % — паразитный.

Набор был разделен в пропорциях 20/80, 40/60, 60/40 и 80/20, где верхняя часть пропорции показывает значение выборки для обучения, а нижняя — для тестирования.

При обучении алгоритмов классификации использовались стандартные настройки Weka. Соответствие кортежей тестируемой выборки одному из классов опасных или безопасных состояний определялось выражениями (2) и (3).

Обучающие выборки формировались искусственно. Изначально предполагалась несбалансированность структуры, для каждого классификатора отдельные события представлены с различной частотой. Моделировалась типовая ситуация, когда классификатор может показывать хорошие результаты на обучающей выборке, но показатели качества падают при работе с тестовыми (в моделируемом случае) или реальными данными, т. е. появляются неучтенные признаки и обучающее множество не полностью отражает генеральную совокупность.

Оценка классификаторов производилась на основе площади под ROC-кривой (AUC) [26] для тестового множества (рисунок).

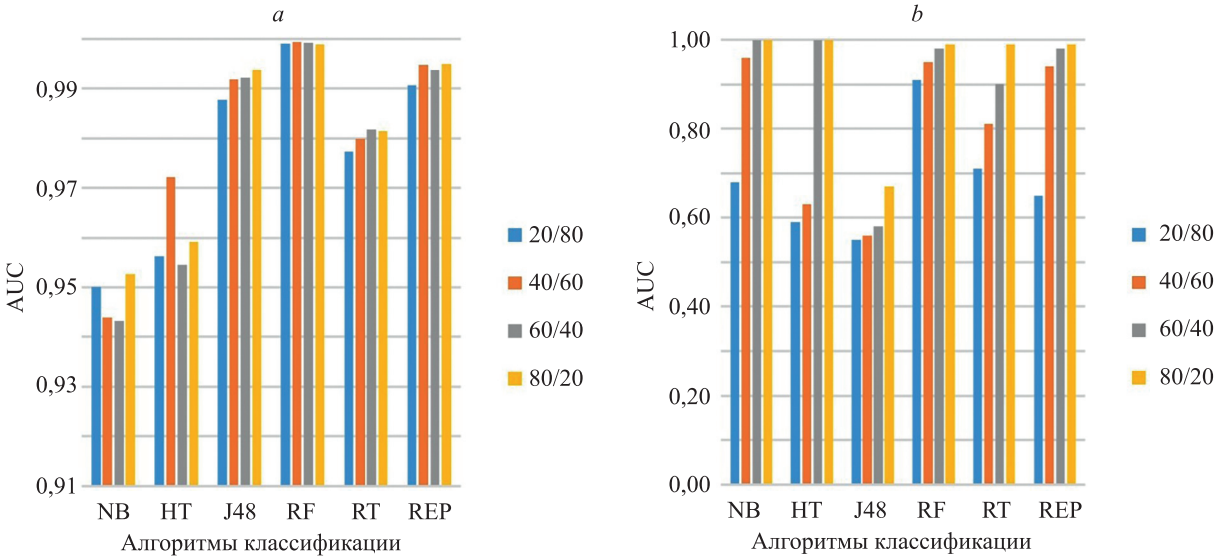

Рисунок. Оценка качества классификаторов по площади под ROC-кривой при сбалансированной (*a*) и несбалансированной (*b*) обучающей выборке в соотношениях 20/80, 40/60, 60/40, 80/20

Figure. Assessment of the quality of classifiers by the area under the ROC-curve with the balanced (*a*) and unbalanced (*b*) training sample in ratios 20/80, 40/60, 60/40, 80/20

На рисунке показан объем выборки экспериментальных данных из датасета и качество алгоритмов для бинарной классификации. Гистограмма (рисунок, *b*) дает возможность определить «слабые» и «сильные» классификаторы, которые возникли после применения несбалансированных обучающих выборок в результате неполноты признаков, неправильного определения зависимостей и других.

Оценка классификаторов осуществлялась на основе значений параметров Accuracy, Precision, Recall, F-мера.

Полученные экспериментальные значения представлены в табл. 1.

Во второй части эксперимента реализован ансамбль классификаторов (5). Поступающие данные одновременно обрабатывались всеми алгоритмами и на основании выражений (4) и (5) определялся класс подмножества $C_j \in C$ который сравнивался с заранее размеченным тестовым множеством.

Таблица 1. Accuracy, Precision, Recall и F-мера для отдельных классификаторов, %

Table 1. Accuracy, Precision, Recall and F-Measure for Individual Classifiers, %

Классификатор	Значения пропорций обучающей к тестирующей выборке			
	20/80	40/60	60/40	80/20
Accuracy				
Naïve Bayes	72,1	96,9	99,9	99,9
Hoeffding Tree	64,2	70,8	99,9	99,9
J48	58,8	59,9	63,6	71,7
Random Forest	94,5	96,8	99,7	99,9
Random Tree	81,6	86,7	96,3	99,9
REP Tree	62,1	95,8	99,9	99,9
Precision				
Naïve Bayes	73,6	98,9	99,9	99,9
Hoeffding Tree	70,7	78,0	99,9	99,9
J48	65,1	66,4	70,4	79,4
Random Forest	95,6	98,0	99,9	99,9
Random Tree	86,0	91,4	99,1	99,9
REP Tree	68,3	98,4	99,9	99,9
Recall				
Naïve Bayes	77,7	96,0	99,9	99,9
Hoeffding Tree	71,1	75,5	99,9	99,9
J48	64,9	65,8	68,5	74,3
Random Forest	95,1	96,7	99,7	99,9
Random Tree	83,3	86,8	95,1	99,9
REP Tree	69,0	95,3	99,9	99,9
F-мера				
Naïve Bayes	75,6	97,4	99,9	99,9
Hoeffding Tree	70,9	76,7	99,9	99,9
J48	65,0	66,1	69,5	76,7
Random Forest	95,3	97,4	99,8	99,9
Random Tree	84,6	89,0	97,1	99,9
REP Tree	68,6	96,8	99,9	99,9

Таблица 2. Результаты применения ансамбля классификаторов

Table 2. Results of applying Bagging of Classifiers

Параметры	Соотношения обучающей к тестовой выборке			
	20/80	40/60	60/40	80/20
	для ансамбля классификаторов			
AUC	0,8947	0,9696	0,9999	0,9999
Accuracy, %	95,0	99,7	99,8	99,9
Precision, %	96,3	99,9	99,9	99,9
Recall, %	96,2	99,6	99,8	99,9
F-мера, %	96,2	99,8	99,9	99,9

Для анализа ансамбля классификаторов выбраны AUC и показатель общей точности последовательности классификаторов. Полученные результаты после применения ансамбля приведены в табл. 2.

Результаты тестирования открытого набора данных NSL-KDD классификаторами машинного обучения, реализованными в приложении Weka, с применением бэггинга, показали точность более 99 %, даже на сравнительно небольшой обучающей выборке.

Основное преимущество предложенного решения состоит в том, что для достижения результатов, сопоставимых с другими методами [19, 20, 23] достаточна относительно небольшая выборка. В отличие от традиционных подходов формирования обучающего множества для алгоритмов создаются разбалансированные обучающие выборки, что дает возможность достичь «разнообразия» классификаторов [18, 21, 25, 26]. Классифицирующие алгоритмы настраиваются для разных видов событий и аномалий, а полученный эффект разброса ответов сглаживается ансамблем. Каждый классификатор «специализируется» на определенной части событий, что позволяет адаптировать его к различным условиям функционирования сетевых сегментов.

Среди недостатков предложенного подхода необходимо отметить чувствительность классифицирующих алгоритмов к смещению ответов. Необходимо заранее анализировать данные, признаковое пространство и классификаторы на предмет возможности возникновения этого эффекта. В случае его сильного влияния результаты ансамбля могут значительно снижаться.

Заключение

Возрастающий объем сетевого трафика обуславливает необходимость развития моделей, методов его анализа на предмет выявления деструктивных воздействий и аномальных ситуаций. Постоянно увеличивается количество новых видов атак, вредоносных сайтов, методов внедрения несанкционированного программного обеспечения. В связи с этим приходится анализировать большое число параметров сетевого информационного обмена.

Одним из основных показателей систем мониторинга является точность идентификации состояния. В работе предложен метод на основе бэггинга классификаторов по выявлению аномальных ситуаций в сетевом трафике. Учитывая большое количество обрабатываемых показателей, предлагаемый подход позволяет получать приемлемые по точности результаты, сглаживая потенциальные ошибки разнородных классификаторов. Разнообразие классификаторов достигается использованием несбалансированных обучающих выборок. В случае применении ансамбля, по мере увеличения объема обучающей выборки, наблюдается рост точности.

Метод имеет ограничение, связанное с возможным проявлением эффекта смещения ответов классифицирующими алгоритмами. В связи с этим перед использованием возникает необходимость дополнительного исследования данных и классификаторов.

Достоинством является возможность его масштабирования и комбинирования путем добавления новых классифицирующих алгоритмов с учетом параметров сетевого трафика в различных сегментах сети.

Литература

1. Khan S., Yairi T. A review on the application of deep learning in system health management // Mechanical Systems and Signal Processing. 2018. V. 107. P. 241–265. doi: 10.1016/j.ymssp.2017.11.024
2. Salehi H., Burgueño R. Emerging artificial intelligence methods in structural engineering // Engineering Structures. 2018. V. 171. P. 170–189. doi: 10.1016/j.engstruct.2018.05.084
3. Gers F.A., Schmidhuber J., Cummins F. Learning to forget: Continual prediction with LSTM // Neural Computation. 2000. V. 12. N 10. P. 2451–2471. doi: 10.1162/089976600300015015
4. Gokhale A., McDonals M.P., Drager S., McKeever W. A cyber physical systems perspective on the real-time and reliable dissemination of information in intelligent transportation systems // Network Protocols and Algorithms. 2010. V. 2. N 3. P. 116–136. doi: 10.5296/npa.v2i3.480
5. Yuan K., Ling Q., Yin W. On the convergence of decentralized gradient descent // SIAM Journal on Optimization. 2016. V. 26. N 3. P. 1835–1854. doi: 10.1137/130943170
6. Kwon D.W., Ko K., Vannucci M., Reddy A.L.N., Kim S. Wavelet methods for the detection of anomalies and their application to network traffic analysis // Quality and Reliability Engineering International. 2006. V. 22. N 8. P. 953–969. doi: 10.1002/qre.781
7. Семенов В.В., Лебедев И.С., Сухопаров М.Е. Подход к классификации состояния информационной безопасности элементов киберфизических систем с использованием побочного электромагнитного излучения // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 1. С. 98–105. doi: 10.17586/2226-1494-2018-18-1-98-105
8. Ahlgren B., Hidell M., Ngai E. Internet of things for smart cities: interoperability and open data // IEEE Internet Computing. 2016. V. 20. N 6. P. 52–56. doi: 10.1109/MIC.2016.124
9. Genkin D., Shamir A., Tromer E. Acoustic cryptanalysis // Journal of Cryptology. 2017. V. 30. N 2. P. 392–443. doi: 10.1007/s00145-015-9224-2
10. Semenov V.V., Lebedev I.S., Sukhoparov M.E., Salakhutdinova K.I. Application of an autonomous object behavior model to classify the cybersecurity state // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2019. V. 11660. P. 104–112. doi: 10.1007/978-3-030-30859-9_9
11. Palacios A., Sanchez L., Couso I. Combining Adaboost with preprocessing algorithms for extracting fuzzy rules from low quality data in possibly imbalanced problems // International Journal of Uncertainty, Fuzziness and Knowledge-based Systems. 2012. V. 20. Suppl. 2. P. 51–71. doi: 10.1142/S0218488512400156
12. Ashibani Y., Mahmoud Q.H. Cyber physical systems security: analysis, challenges and solutions // Computer & Security. 2017. V. 68. P. 81–97. doi: 10.1016/j.cose.2017.04.005
13. Jin J., Gubbi J., Marusic S., Palaniswami M. An information framework for creating a smart city through internet of things // IEEE Internet of Things Journal. 2014. V. 1. N 2. P. 112–121. doi: 10.1109/JIOT.2013.2296516
14. Сухопаров М.Е., Семенов В.В., Салахутдинова К.И., Лебедев И.С. Выявление аномального функционирования устройств Индустрии 4.0 на основе поведенческих паттернов // Проблемы информационной безопасности. Компьютерные системы. 2020. № 1. С. 96–102.

References

1. Khan S., Yairi T. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 2018, vol. 107, pp. 241–265. doi: 10.1016/j.ymssp.2017.11.024
2. Salehi H., Burgueño R. Emerging artificial intelligence methods in structural engineering. *Engineering Structures*, 2018, vol. 171, pp. 170–189. doi: 10.1016/j.engstruct.2018.05.084
3. Gers F.A., Schmidhuber J., Cummins F. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 2000, vol. 12, no. 10, pp. 2451–2471. doi: 10.1162/089976600300015015
4. Gokhale A., McDonals M.P., Drager S., McKeever W. A cyber physical systems perspective on the real-time and reliable dissemination of information in intelligent transportation systems. *Network Protocols and Algorithms*, 2010, vol. 2, no. 3, pp. 116–136. doi: 10.5296/npa.v2i3.480
5. Yuan K., Ling Q., Yin W. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 2016, vol. 26, no. 3, pp. 1835–1854. doi: 10.1137/130943170
6. Kwon D.W., Ko K., Vannucci M., Reddy A.L.N., Kim S. Wavelet methods for the detection of anomalies and their application to network traffic analysis. *Quality and Reliability Engineering International*, 2006, vol. 22, no. 8, pp. 953–969. doi: 10.1002/qre.781
7. Semenov V.V., Lebedev I.S., Sukhoparov M.E. Approach to classification of the information security state of elements for cyberphysical systems by applying side electromagnetic radiation. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 1, pp. 98–105. (in Russian). doi: 10.17586/2226-1494-2018-18-1-98-105
8. Ahlgren B., Hidell M., Ngai E. Internet of things for smart cities: interoperability and open data. *IEEE Internet Computing*, 2016, vol. 20, no. 6, pp. 52–56. doi: 10.1109/MIC.2016.124
9. Genkin D., Shamir A., Tromer E. Acoustic cryptanalysis. *Journal of Cryptology*, 2017, vol. 30, no. 2, pp. 392–443. doi: 10.1007/s00145-015-9224-2
10. Semenov V.V., Lebedev I.S., Sukhoparov M.E., Salakhutdinova K.I. Application of an autonomous object behavior model to classify the cybersecurity state. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11660, pp. 104–112. doi: 10.1007/978-3-030-30859-9_9
11. Palacios A., Sanchez L., Couso I. Combining Adaboost with preprocessing algorithms for extracting fuzzy rules from low quality data in possibly imbalanced problems. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 2012, vol. 20, suppl. 2, pp. 51–71. doi: 10.1142/S0218488512400156
12. Ashibani Y., Mahmoud Q.H. Cyber physical systems security: analysis, challenges and solutions. *Computer & Security*, 2017, vol. 68, pp. 81–97. doi: 10.1016/j.cose.2017.04.005
13. Jin J., Gubbi J., Marusic S., Palaniswami M. An information framework for creating a smart city through internet of things. *IEEE Internet of Things Journal*, 2014, vol. 1, no. 2, pp. 112–121. doi: 10.1109/JIOT.2013.2296516
14. Sukhoparov M.E., Semenov V.V., Salakhutdinova K.I., Lebedev I.S. Identification of anomalous functioning of Industry 4.0 devices based on behavioral patterns. *Information Security Problems. Computer Systems*, 2020, no. 1, pp. 96–102. (in Russian)
15. Semenov V., Lebedev I., Sukhoparov M. Identification of the state of individual elements of cyber-physical systems based on external

15. Семенов В.В., Лебедев И.С., Сухопаров М.Е. Идентификация состояния отдельных элементов киберфизических систем на основе внешних поведенческих характеристик // Прикладная информатика. 2018. Т. 13. № 5(77). С. 72–83.
16. Сухопаров М.Е., Лебедев И.С. Идентификация состояния информационной безопасности устройств интернета вещей в информационно-телекоммуникационных системах // Системы управления, связи и безопасности. 2020. № 3. С. 252–268. doi: 10.24411/2410-9916-2020-10310
17. Sukhoparov M.E., Lebedev I.S., Garanin A.V. Application of classifier sequences in the task of state analysis of Internet of Things devices // Информатика, телекоммуникации и управление = Computing, Telecommunications and Control. 2020. Т. 13. № 3. С. 44–54. doi: 10.18721/JCSTCS.13304
18. Ingre B., Yadav A. Performance Analysis of NSL-KDD dataset using ANN // Proc. 4th International Conference on Signal Processing and Communication Engineering Systems (SPACES). 2015. P. 92–96. doi: 10.1109/SPACES.2015.7058223
19. Dhanabal L., Shantharajah Dr. S.P. A Study on NSL-KDD dataset for intrusion detection system based on classification algorithms // International Journal of Advanced Research in Computer and Communication Engineering. 2015. V. 4. N 6. P. 446–452. doi: 10.17148/IJARCCE.2015.4696
20. Воронцов К.В. Лекции по алгоритмическим композициям [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/images/0/0d/Voron-ML-Compositions.pdf (дата обращения: 03.12.2020).
21. Дьяконов А. Методы решения задач классификации с категориальными признаками // Прикладная математика и информатика. Труды факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова. 2014. № 46. С. 103–127.
22. Zhou Z.-H. Ensemble Methods: Foundations and Algorithms. New York: CRC Press, 2012. 222 p.
23. Yu Y., Zhou Z.-H., Ting K.M. Cocktail ensemble for regression // Proc. 7th IEEE International Conference on Data Mining (ICDM). 2007. P. 721–726. doi: 10.1109/ICDM.2007.60
24. Zhou Z.-H., Feng J. Deep forest // National Science Review. 2019. V. 6. N 1. P. 74–86. doi: 10.1093/nsr/nwy108
25. Pedersen T. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation // NAACL 2000: Proc. of the 1st North American chapter of the Association for Computational Linguistics Conference. 2000. P. 63–69.
26. Кафтанников И.Л., Парасич А.В. Проблемы формирования обучающей выборки в задачах машинного обучения // Вестник ЮУрГУ. Серия: Компьютерные технологии, управление, радиоэлектроника. 2016. Т. 16. № 3. С. 15–24. doi: 10.14529/ctcr160302
27. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. 2006. V. 27. N 8. P. 861–874. doi: 10.1016/j.patrec.2005.10.010

behavioral characteristics. *Journal of Applied Informatics*, 2018, vol. 13, no. 5(77), pp. 72–83. (in Russian)
16. Sukhoparov M.E., Lebedev I.S. Identification the information security status for the internet of things devices in information and telecommunication systems. *Systems of Control, Communication and Security*, 2020, no. 3, pp. 252–268. (in Russian). doi: 10.24411/2410-9916-2020-10310
17. Sukhoparov M.E., Lebedev I.S., Garanin A.V. Application of classifier sequences in the task of state analysis of Internet of Things devices. *Computing, Telecommunications and Control*, 2020, vol. 13, no. 3, pp. 44–54. doi: 10.18721/JCSTCS.13304
18. Ingre B., Yadav A. Performance Analysis of NSL-KDD dataset using ANN. *Proc. 4th International Conference on Signal Processing and Communication Engineering Systems (SPACES)*, 2015, pp. 92–96. doi: 10.1109/SPACES.2015.7058223
19. Dhanabal L., Shantharajah Dr. S.P. A Study on NSL-KDD dataset for intrusion detection system based on classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 2015, vol. 4, no. 6, pp. 446–452. doi: 10.17148/IJARCCE.2015.4696
20. Vorontcov K.V. *Lectures on algorithmic compositions*. Available at: http://www.machinelearning.ru/wiki/images/0/0d/Voron-ML-Compositions.pdf (accessed: 03.12.2020). (in Russian)
21. D'yakonov A.G. Solution methods for classification problems with categorical attributes. *Computational Mathematics and Modeling*, 2015, vol. 26, no. 3, pp. 408–428. doi: 10.1007/s10598-015-9281-2
22. Zhou Z.-H. *Ensemble Methods: Foundations and Algorithms*. New York, CRC Press, 2012, 222 p.
23. Yu Y., Zhou Z.-H., Ting K.M. Cocktail ensemble for regression. *Proc. 7th IEEE International Conference on Data Mining (ICDM)*, 2007, pp. 721–726. doi: 10.1109/ICDM.2007.60
24. Zhou Z.-H., Feng J. Deep forest. *National Science Review*, 2019, vol. 6, no. 1, pp. 74–86. doi: 10.1093/nsr/nwy108
25. Pedersen T. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. *NAACL 2000: Proc. of the 1st North American chapter of the Association for Computational Linguistics Conference*, 2000, pp. 63–69.
26. Kaftannikov I.L., Parasich A.V. Problems of training set's formation in machine learning tasks. *Bulletin of the South Ural State University. Series Computer Technology, Aotimatic Control, Radio Electronics*, 2016, vol. 16, no. 3, pp. 15–24. (in Russian). doi: 10.14529/ctcr160302
27. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006, vol. 27, no. 8, pp. 861–874. doi: 10.1016/j.patrec.2005.10.010

Авторы

Рзаев Бабыр Темирбекулы — магистр, докторант, Казахский агротехнический университет им. С. Сейфуллина, Нур-Султан, 010000, Республика Казахстан, https://orcid.org/0000-0002-9671-650X, pathinchaos@gmail.com

Лебедев Илья Сергеевич — доктор технических наук, профессор, заведующий лабораторией, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 56321781100, https://orcid.org/0000-0001-6753-2181, isl_box@mail.ru

Authors

Babyr T. Rzayev — M.Sc., Postgraduate, Saken Seifullin Kazakh Agrotechnical University, Nur-Sultan, 010000, Kazakhstan, https://orcid.org/0000-0002-9671-650X, pathinchaos@gmail.com

Ilya S. Lebedev — D.Sc., Professor, Laboratory Head, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 56321781100, https://orcid.org/0000-0001-6753-2181, isl_box@mail.ru

Статья поступила в редакцию 20.12.2020
Одобрена после рецензирования 19.02.2021
Принята к печати 16.03.2021

Received 20.12.2020
Approved after reviewing 19.02.2021
Accepted 16.03.2021