УДК 004.85 Вестник УрФО № 1(35) / 2020, с. 77–83

Асяев Г. Д., Соколов А. Н.

DOI: 10.14529/secur200109

ОБНАРУЖЕНИЕ ВТОРЖЕНИЙ НА ОСНОВЕ АНАЛИЗА АНОМАЛЬНОГО ПОВЕДЕНИЯ ЛОКАЛЬНОЙ СЕТИ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ

В работе представлены модели процесса обнаружения вторжений, построенные на основе трёх методов машинного обучения: метода деревьев решений, метода ближайших соседей и метода случайного леса. Основной задачей при моделировании является классификация состояний автоматизированной системы управления (АСУ) (аномальное, нормальное). Рассмотрены параметры, влияющие на обнаружение аномального поведения: протокол, сервисные данные, используемые флаги, количество неудачных попыток входа, продолжительность атаки. Для моделирования процесса поиска аномалий выбран набор данных транспортно-сетевого уровня АСУ, состоящий из необработанных дампов TCP/IP в ситуации, когда сеть подверглась множественным атакам. Для каждого соединения TCP/IP фиксировались 3 качественных и 38 количественных признаков, среди которых выделены наиболее важные признаки, влияющие на обучение. Прогнозирование ответа проводилось на контрольной (тестовой) выборке. Основными критериями выбора математической модели для решаемой задачи являлись количество правильно распознанных (accuracy) аномалий, точность (precision) и полнота (recall) ответов. На основании проведенных исследовании был выбран оптимальный алгоритм для обнаружения аномалий.

Ключевые слова: *автоматизированная система управления, обнаружение вторжений, уязвимость, аномалия, метод машинного обучения, метод деревьев решений, метод ближайших соседей, метод случайного леса.*

DETECTION OF INVASION ON THE BASIS OF ANALYSIS OF ANOMALOUS BEHAVIOR OF A LOCAL NETWORK USING MACHINE-LEARNING ALGORITHMS WITH A TEACHER

The paper presents models of the intrusion detection process based on three machine learning methods: the decision tree method, the nearest neighbor method and the random forest method. The main task in modeling is to classify the ACS states (abnormal, normal). Parameters affecting the detection of anomalous behavior are considered: protocol, service data, flags used, number of unsuccessful attempts to enter, duration of the attack. To simulate the process of anomaly detection, the data set of the transport and network level of the control system, consisting of raw TCP/IP dumps in a situation where the network has been subjected to multiple attacks, was selected. For each TCP/IP connection, 3 qualitative and 38 quantitative features were recorded, among which the most important features affecting the learning were highlighted. The response was predicted in a control (test) sample. The main criteria for choosing a mathematical model for the task were the number of correctly recognized (accuracy) anomalies, accuracy (precision) and completeness (recall) of answers. The optimal algorithm for detection of anomalies was chosen on the basis of the conducted research.

Keywords: *automated control system, intrusion detection, vulnerability, anomaly, machine learning method, decision tree method, closest neighbour method, random forest method..*

Автоматизированные системы управления (АСУ) широко используется во многих отраслях производственной деятельности. Через их элементы управления проходят большие объемы данных (big data), основной угрозой для которых является вмешательство террористических, экстремистских и враждебно настроенных групп в управление автоматизированными системами, в том числе с целью вывода их из строя [1]. Количество примеров подобных атак (Stuxnet, Crouching Yeti, BlackEnergy и т.п.) ежегодно растет вследствие большого числа уязвимостей у эксплуатируемых систем. Влияние уязвимости на вероятность осуществления атаки тем выше, чем [1]:

• большее число узлов, реализующих функцию, подвержено уязвимости;

• большее число функций обслуживается уязвимым программным обеспечением.

Для обеспечения безопасности информации АСУ необходимо обеспечение непрерывного контроля трафика всех взаимодействий системы. Постоянный анализ этих данных позволяет своевременно выявлять аномальное поведение системы, связанное с её некорректным функционированием. Наиболее критичной является задача сохранения способности АСУ к корректному функционированию в условиях деструктивных информационных воздействий. Успешная реализация кибератак на такие системы может повлечь за собой негативные финансовые последствия, экологические катастрофы или даже привести к гибели людей. Поэтому важными особенностями АСУ являются [2]:

• непрерывный режим работы: остановка работы АСУ, как правило, либо невозможна, либо влечет за собой значительные финансовые потери;

• в зависимости от особенностей технологического процесса различные элементы автоматизированной информационной системы вносят различный вклад по степени критичности возможного ущерба.

Для поддержания системы информационной безопасности АСУ на требуемом уровне, необходима регулярная установка критических обновлений, направленных на исправление уязвимостей. Тем не менее, в силу вышеуказанных особенностей АСУ, это не всегда рационально [3].

сировались 3 качественных и 38 количественных признаков. При анализе данных использованы следующие признаки (см. рис. 1):

• duration – продолжительность атаки;

• protocol_type – используемый протокол;

• service – сервисные данные;

• flag – используемые флаги;

• num_failed_logins – количество неудачных попыток входа в систему;

• logged_in – количество вхождений в систему;

• class – критерий, характеризующий поведение системы как нормальное или аномальное.

На рис. 2 приведена диаграмма, показывающая соотношение нормального и ано-

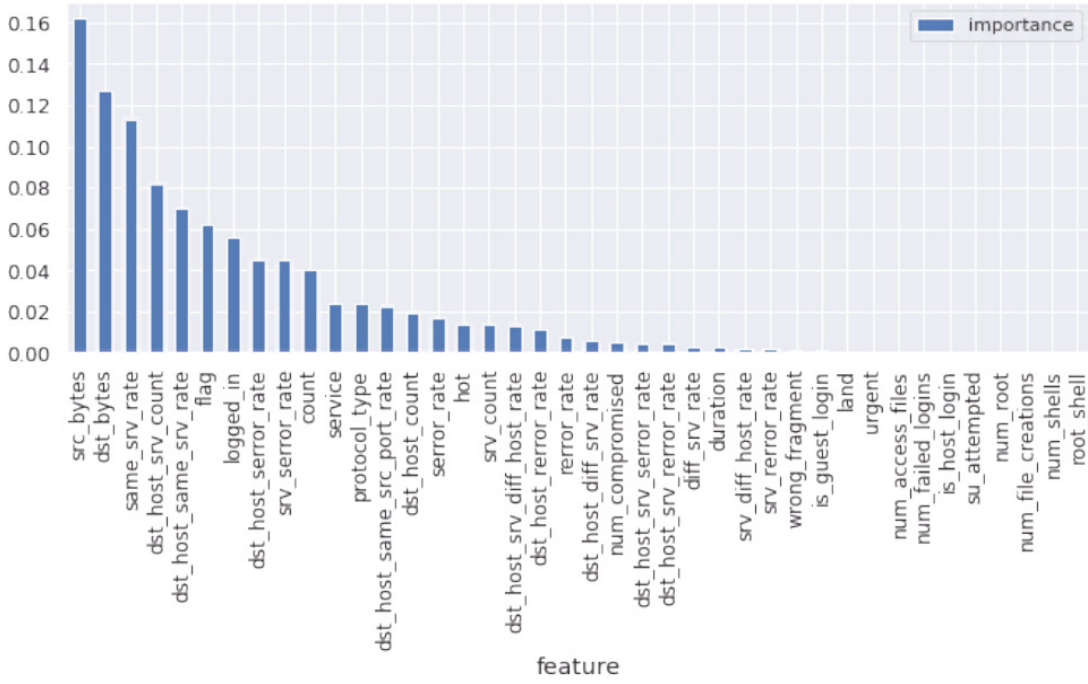
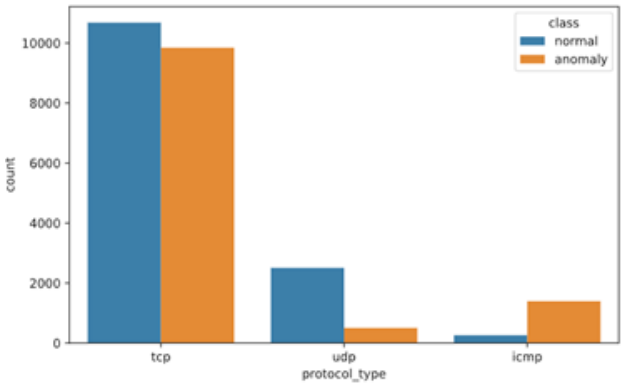


Рис. 1. Диаграмма значимости признаков по степени их влияния на обучение

Для моделирования процесса поиска аномалий выбран набор данных транспортно-сетевого уровня АСУ, состоящий из необработанных дампов TCP/IP в ситуации, когда сеть подверглась множественным атакам. Каждое соединение состоит из последовательности TCP-пакетов, начинающихся и заканчивающихся в моменты времени, в промежутке между которыми данные по определенному протоколу передаются на IP-адрес источника, а с него – на целевой IP-адрес. Кроме того, каждое соединение помечается как нормальное или как атака определенного типа. Размер каждой записи соединения – около 100 байт. Для каждого соединения TCP/IP фик-

мального состояния признака class для трёх типов протоколов (TCP, UDP и ICMP). Из рисунка видно, что наибольшее общее число аномальных состояний возникает при передаче данных с использованием протокола TCP (9845 или 48,0 %). Однако можно заметить, что злоумышленник активно использует протокол межсетевых управляющих сообщений ICMP, так как количество срабатываний счетчика аномалий примерно в 5 раз выше по сравнению с нормальным режимом работы (1394 или 84,2 %) [4]. Наименьшее количество аномалий наблюдается при использовании протокола UDP (504 или 16,7 %).

Диаграммы распределений аномального



class	anomaly	normal	All
protocol_type			
icmp	1394	261	1655
tcp	9845	10681	20526
udp	504	2507	3011
All	11743	13449	25192

Рис. 2. Соотношение нормального и аномального состояния признака class для протоколов TCP, UDP и ICMP

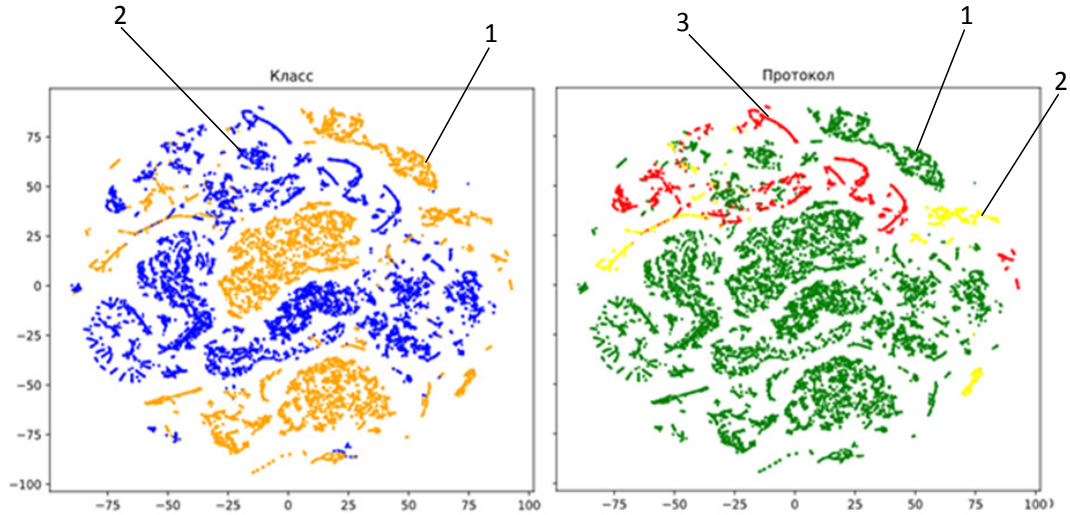


Рис. 3. Распределения аномального (оранжевый, 1) и нормального (синий, 2) состояний системы в зависимости от используемого протокола: TCP (зеленый, 1), UDP (желтый, 2) и ICMP (красный, 3)

и нормального состояний системы в зависимости от типов протоколов приведены на рис. 3.

Моделирование процесса поиска аномалий проведено с использованием трёх методов машинного обучения: метода деревьев решений, метода ближайших соседей и метода случайного леса. Основной задачей при моделировании является классификация состояния системы (аномальное, нормальное). Моделирование проводилось в среде Jupyter Notebook с использованием библиотек scikit-learn, pandas, numpy. Для обучения математических моделей категориальные признаки были закодированы с использованием метода «one-hot encoding».

Метод деревьев решений. Разделим обучающуюся выборку в соотношении 70/30, предварительно перемешав строки для оптимального обучения на соответствующей выборке. Размер обучающейся выборки равен 17634, при этом количество уникальных признаков равно 39. Важным параметром для метода деревьев решений является глубина дерева. Уменьшение глубины дерева приводит к недообучению, в то время как увеличение глубины может переобучить сеть, а математическая модель будет некорректно работать с новыми данными [5].

Значение кросс-валидации выберем равным 5, а глубину дерева 4. Точность распознанных аномалий при этом составила 97,72 %. Лучше всего эту модель удалось обучить при глубине дерева, равной 10, и параметре кросс-валидации, равному 10. Количество признаков, по которым произошло наилучшее разбиение в дереве, равно 25. Ниже представлены оптимальные параметры дерева, применяемые в обучении этой модели:

criterion = '**gini**', splitter = '**best**', max_depth = **10**, min_samples_split = **2**, min_samples_leaf = **1**, min_weight_fraction_leaf = **0.0**, max_features = **25**, random_state = **17**, max_leaf_nodes = **None**, min_impurity_decrease = **0.0**, min_impurity_split = **None**, class_weight = **None**, presort = **False**.

Выборка была разделена на обучающую и тестовую. Precision = 92,22%, recall =91,08%. Доля распознанных аномалий на обучающей выборке составила 99,56 %, что свидетельствует о хорошей степени обученности модели (см. рис. 4).

На вход обученной модели подавалась тестовая выборка. Доля правильных ответов на тестовой выборке составило 99,02 %. Таким образом, представленная математическая модель достаточно хорошо распознает аномалии.

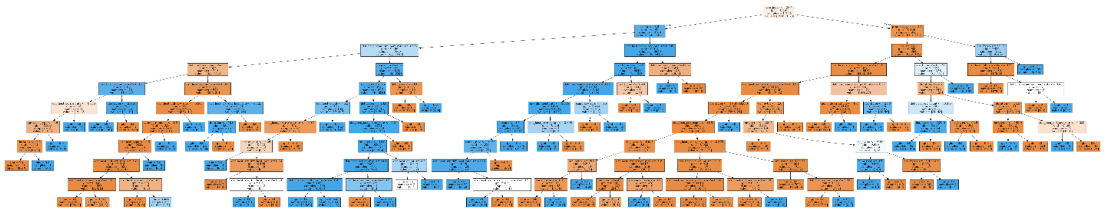


Рис. 4. Дерево решений обнаружения аномалий

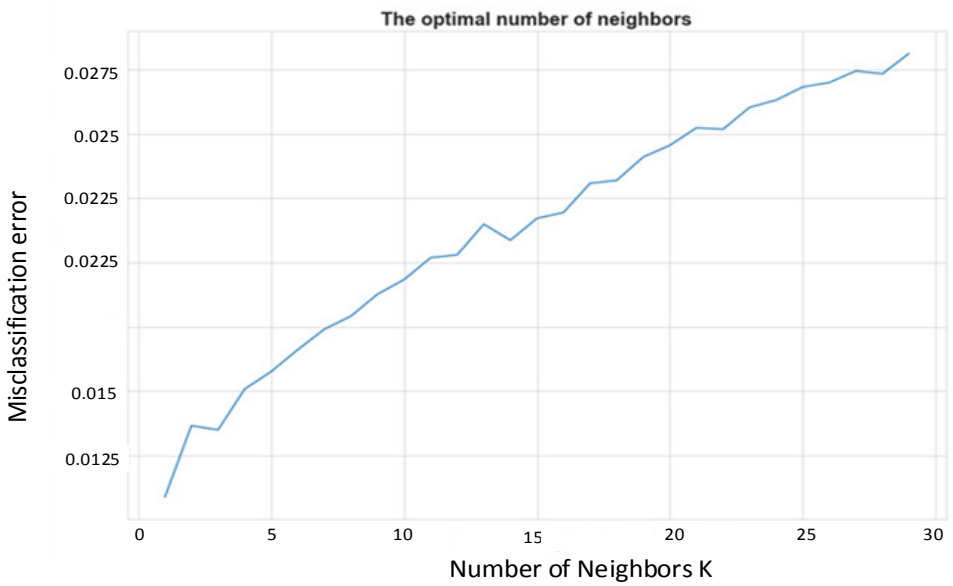


Рис. 5. Зависимость числа ближайших соседей от среднеквадратичной ошибки

Метод ближайших соседей. Данный метод относится к методам классификации без учителя [6]. Данные группируются по схожему признаку на основе рассчитанных весов. Для обучения модели выбрано значение кросс-валидации, равное 10, как и в методе деревьев решений. Максимальную долю правильных ответов удалось получить при числе ближайших соседей, равным 1. Доля правильных ответов составила 98,91 %. Ниже представлены оптимальные параметры, применяемые в обучении данной модели методом ближайших соседей (см. рис. 5):

n_neighbors = **1**, weights = '**uniform**', algorithm = '**auto**', leaf_size = **30**, p = **2**, metric = '**minkowski**', metric_params = **None**, n_jobs = **None**.

На тестовой выборке количество правильно распознанных аномалий составило 99,03 %, что соизмеримо с результатами, полученными с использованием метода деревьев решений. Precision = 93,92%, recall = 95,78%.

Метод случайного леса (Random forest). Этот метод построен на совокупности деревьев решений, прогноз которых усредняется. Такой подход имеет очевидные преимущества, связанные с повышенной точностью прогноза и минимизация процесса обучения [7]. Однако, при этом он требует значительных вычислительных мощностей. Максимальную точность прогноза удалось получить при максимальной глубине дерева, равной 10, и количестве признаков, равным 11. Точность прогноза составила 99,64 %. Precision = 98,29%, recall =98,68%.

На тестовой выборке количество правильно распознанных аномалий составило 99,69 %, что является лучшим результатом среди представленных:

RandomForestClassifier(n_estimators =

'**warn**', criterion = '**gini**', max_depth = **10**, min_samples_split = **2**, min_samples_leaf = **1**, min_weight_fraction_leaf = **0.0**, max_features = **11**, max_leaf_nodes = **None**, min_impurity_decrease = **0.0**, min_impurity_split = **None**, bootstrap = **True**, oob_score = **False**, n_jobs = **None**, random_state = **None**, verbose = **0**, warm_start = **False**, class_weight = **None**).

Специфика работы АСУ, как правило такова, что остановка её функционирования с целью обновления систем безопасности [8], обновления сигнатур антивирусных баз данных, может повлечь за собой значительные финансовые потери. Одним из решений, минимизирующим необходимость такой остановки, является анализ сетевого трафика с использованием методов машинного обучения. Анализ трафика позволяет распознавать аномалии в режиме реального времени, и, на основании этого, обнаруживать уязвимости системы. Обнаруженные уязвимости определяют необходимость установки критических обновлений. Наилучшие результаты среди рассмотренных получены с использованием метода случайного леса (Random forest): количество правильно распознанных аномалий составило 99,69 %, а полнота и точность (многообразие правильно распознанных ответов и точность отнесения к аномалии) составила 98%.

Статья выполнена при поддержке Правительства РФ (Постановление №211 от 16.03.2013 г.), соглашение № 02.A03.21.0011.

Литература

1. Андрей Заикин. Почему защита АСУ ТП сегодня стала критически важной? [Электронный ресурс] //https://www.securitylab.ru/analytics/484730.php. (Дата обращения: 10.03.2020).

2. A. Mansouri, B. Majidi and A. Shamisa, "Anomaly detection in industrial control systems using evolutionary-based optimization of neural networks", Communications on Advanced Computational Science with Applications, vol. 2017, no. 1, pp. 49–55. Available: 10.5899/2017/cacsa-00074.

3. А.Е. Баринов, С.В. Скурлаев, А.Н. Соколов. "Методика оценки рисков, вызванных уязвимостями в программном обеспечении автоматизированных систем управления технологическими процессами", Вестник УрФО. Безопасность в информационной сфере., № 3 (25), с. 34–42, 2017.

4. C. Feng, T. Li, D. Chana, "Multi-level anomaly detection in industrial control systems via package signatures and lstm networks", Dependable Systems and Networks (DSN), 47th Annual IEEE/IFIP International Conference on. – IEEE, vol. 47, pp. 261–272, 2017.

5. А.А. Бранитский, И.В. Котенко, "Анализ и классификация методов обнаружения сетевых атак", Труды СПИИРАС, № 45, с. 207–244, 2016.

6. S. N. Shirazi, "Evaluation of anomaly detection techniques for scada communication resilience", Resilience Week (RWS), IEEE, pp. 140–145, 2016.

7. М.С. Пырьев, А.С. Коллеров, "Средства анализа сетевого трафика локальной вычислительной сети в ретроспективе", Вестник УрФО. Безопасность в информационной сфере, (4(34)), с. 58–62, 2019.

8. M. Chandrashekar, Y. Lee and D. Medhi, "Real-time network anomaly detection system using machine learning", 11th International Conference on the Design of Reliable Communication Networks (DRCN), Kansas City, MO, vol. 11, pp. 267–270, 2015.

References

1. Andrey Zaikin. Why is the protection of process control systems now critical? [Electronic resource] // https://www.securitylab.ru/analytics/484730.php. (Date of treatment: 10.03.2020).

3. A. E. Barinov, S. V. Skurlaev, A. N. Sokolov, "Methodology for assessing the risks caused by vulnerabilities in the software of automated process control systems", Bulletin of the Urals Federal District. Security in the information field., vol. 3(25), pp. 34-42, 2017.

5. A.A. Branitsky, I.V. Kotenko, "Analysis and classification of network attack detection methods", Tr. SPIIRAS, vol. 45, pp. 207–244, 2016.

7. M.S. Pyryev, A.S. Kollerov, "A retrospective analysis of the network traffic of a local computer network", Proceedings of the XVIII All-Russian Scientific and Practical Conference of Students, Graduate Students and Young Scientists "Information Space Security". – Magnitogorsk, vol. 18, pp.. 302–306, 2019.

АСЯЕВ Григорий Дмитриевич, аспирант кафедры защиты информации ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)». Россия, 454080 Челябинск, проспект Ленина, 76. E-mail: asiaevgd@susu.ru.

СОКОЛОВ Александр Николаевич, кандидат технических наук, доцент, заведующий кафедрой защиты информации ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)». Россия, 454080 Челябинск, проспект Ленина, 76. E-mail: sokolovan@susu.ru.

ASYAEV Grigorii, Postgraduate Student, Department of Information Security, South Ural State University (National Research University). 76, Lenin prospekt, Chelyabinsk, Russia, 454080. E-mail: asiaevgd@susu.ru.

SOKOLOV Alexander, Ph.D., Associate Professor, Head of Department of Information Security, Federal State Autonomous Educational Institution of Higher Education “South Ural State University (national research university)”. 76, Lenin prospekt, Chelyabinsk, Russia, 454080. E-mail: sokolovan@susu.ru.