ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ
INFORMATION TECHNOLOGY AND TELECOMMUNICATIONS

УДК004.8

DOI: 10.21822/2073-6185-2024-51-1-106-112 Оригинальная статья /Original article

Исследование обнаружения аномалий с использованием Isolation Forest в машинном обучении

А.С. Кечеджиев, О.Л. Цветкова
Донской государственный технический университет
344003, г. Ростов-на-Дону, пл. Гагарина 1, Россия

Резюме. Цель. Исследование посвящено оценке применимости метода Isolation Forest в задаче обнаружения аномалий на данных сетевого трафика, характеризующихся недостаточной разметкой. Основной целью данной работы является оценка эффективности Isolation Forest при ограниченной разметке данных и его потенциала в критически важных областях, таких как кибербезопасность и финансовая аналитика. **Метод**. Исследование включает предварительную обработку данных, обучение модели на тренировочном наборе, а также оценку производительности модели на тестовом наборе с использованием метрик точности, матрицы ошибок и отчета о классификации. Для реализации данного исследования был выбран язык программирования Python и библиотека scikit-learn для реализации Isolation Forest, а также Pandas для работы с данными. **Результат**. Оценка применимости метода Isolation Forest на неструктурированных данных выявила его потенциал в выделении аномальных паттернов без необходимости обширной разметки. Это подтверждает эффективность Isolation Forest в условиях, где доступ к размеченным данным ограничен или отсутствует. **Вывод.** Полученные результаты демонстрируют высокую полноту обнаружения аномалий, несмотря на относительно низкую общую точность, что указывает на важность контекстуальной интерпретации метрик в задаче обнаружения редких событий в данных.

Ключевые слова: искусственный интеллект, машинное обучение, кибербезопасность, киберугрозы, информационная безопасность, Isolation Forest.

Для цитирования: А.С. Кечеджиев, О.Л. Цветкова. Исследование обнаружения аномалий с использованием Isolation Forest в машинном обучении. Вестник Дагестанского государственного технического университета. Технические науки. 2024; 51(1):106-112. DOI:10.21822/2073-6185-2024-51-1-106-112

Anomaly detection research using Isolation Forest in Machine Learning

A. S. Kechedzhiev, O. L. Tsvetkova
Don State Technical University,
1 Gagarin Square, Rostov-on-Don 344003, Russia

Abstract. Objective. The study is devoted to assessing the applicability of the Isolation Forest method in the task of detecting anomalies in network traffic data characterized by insufficient markup. The main purpose of the work is to evaluate the effectiveness of Isolation Forest with limited data markup and its potential in critical areas such as cybersecurity and financial analytics. **Method.** The study includes data preprocessing, training the model on the training set, and evaluating the model's performance on the test set using accuracy metrics, error matrix, and classification report. To implement this research, the Python programming language and the scikit-learn library were chosen to implement the Isolation Forest, as well as Pandas for working with data. **Result.** Evaluating the applicability of the Isolation Forest method on unstructured data revealed its potential for identifying anomalous patterns without the need for extensive labeling. This confirms the effectiveness of Isolation Forest in environments where access to labeled data is limited or absent. **Conclusion.** The results demonstrate high anomaly detection recall despite relatively low overall

accuracy, indicating the importance of contextual interpretation of metrics in the task of detecting rare events in data.

Keywords: artificial intelligence, machine learning, cybersecurity, cyberthreats, information security, Isolation Forest

For citation: A.S. Kechedzhiev, O.L. Tsvetkova. Anomaly detection research using Isolation Forest in Machine Learning. Herald of Daghestan State Technical University. Technical Sciences. 2024; 51(1): 106-112. DOI:10.21822/2073-6185-2024-51-1-106-112

Введение. Современные задачи обработки данных, особенно в контексте обнаружения аномалий, часто сталкиваются с отсутствием подробной разметки или недостаточным объемом размеченных данных для обучения моделей. Недостаточная или отсутствующая разметка данных затрудняет создание точных моделей, способных выявлять аномалии в реальном времени. Эта проблема становится особенно актуальной в областях, где нештатные ситуации могут иметь серьезные последствия, таких как кибербезопасность, финансовая аналитика и производственные процессы.

Основные проблемы заключаются в следующем:

Недостаток размеченных данных. В реальных условиях часто бывает сложно или дорого создать большой объем размеченных данных для обучения моделей обнаружения аномалий. Это ставит под вопрос эффективность и точность моделей, требующих разметку для обучения.

Необходимость оперативного обнаружения аномалий. Во многих областях, таких как кибербезопасность или финансовая аналитика, нештатные ситуации требуют мгновенного реагирования. Оперативное обнаружение и предотвращение аномальных событий становится критически важным.

Точность и надежность обнаружения. Отсутствие разметки или недостаточное количество размеченных данных может привести к недостаточной точности моделей обнаружения аномалий. Важно создать методы, способные точно выделять аномалии при минимальном использовании размеченных данных.

Выявление аномалий представляет собой процесс определения данных, элементов, наблюдений или событий, которые отличаются от большинства образцов в наборе данных. Обнаружение аномалий может указывать на подозрительный сетевой трафик, неисправности измерительных приборов, ошибки измерений или просто данные, которые следует исключить перед последующим анализом [1]. Большинство исследований в области обнаружения аномалий основываются на моделях, которые обучаются на ограниченных выборках [2, 3]. Однако такие модели часто имеют погрешность, так как требуется их переобучение в процессе работы [3]. Некоторые авторы предлагают использованием машинного обучения для того, чтобы иметь возможность поиска аномального поведения пользователей по журналам событий безопасности Windows [4]. Интерес представляет решение такой задачи как динамическая аутентификация пользователей, выполненная по результатам анализа работы с компьютерной мышью [5]. Модифицированные алгоритмы кластерного анализа могут быть применены для обнаружения аномалий на основе машинного обучения [6].

Постановка задачи. Процесс обнаружения аномалий сталкивается с острой проблемой нехватки размеченных данных, что затрудняет точное выявление аномалий и требует методов, способных работать с ограниченной разметкой.

В таких условиях изоляционный лес выделяется как перспективный инструмент. Его эффективность в обнаружении аномалий при минимальном использовании размеченных данных делает его привлекательным выбором.

Метод основан на принципе изоляции аномалий путем случайного разделения, что позволяет оперативно и точно выявлять нештатные ситуации. Кроме того, его низкая чувствительность к выбросам в данных делает его устойчивым к шуму и способным эффективно обрабатывать различные типы аномалий.

Целью исследования является изучение и оценка метода Isolation Forest в контексте обнаружения аномалий в данных, где отсутствует детальная разметка. Основная задача состоит в проверке применимости этого метода в условиях неструктурированных данных, где доступ к полной разметке ограничен или отсутствует. Также необходимо оценить эффективность Isolation Forest в определении аномальных паттернов в данных без необходимости наличия большого объема размеченных образцов.

Это исследование будет осуществляться путем изучения способностей метода Isolation Forest на различных типах данных, включая как структурированные, так и неструктурированные данные разного формата и представления. Основное внимание уделяется оценке производительности метода на данных реальных сценариев применения, таких как кибербезопасность, финансовая аналитика или обнаружение нештатных ситуаций, где точное и оперативное обнаружение аномалий имеет важное значение.

Для достижения цели исследования будет проведен обширный анализ результатов, полученных при применении Isolation Forest к различным наборам данных. Это включает в себя оценку производительности модели через различные метрики, такие как точность, полнота, F-мера и другие, для более глубокого понимания способности модели точно выделять аномалии в данных. Основная задача состоит в том, чтобы понять, насколько успешно метод Isolation Forest может быть применен в условиях ограниченной разметки данных и оценить его потенциал для использования в различных областях, где обнаружение аномалий является важным аспектом анализа данных.

Важным аспектом исследования является оценка возможности обобщения результатов на другие области и типы данных для понимания того, как полученные выводы могут быть применимы в различных контекстах и сценариях.

Методы исследования. Для обучения взят набор данных, который содержит сетевой трафик, собранный в контролируемой лабораторной среде. Этот набор данных был создан для обнаружения вредоносного сетевого трафика (сетевых атак) с использованием методов машинного обучения. Он содержит информацию о различных видах сетевых атак, таких как DoS, атаки на переполнение буфера, различные типы сканирования и другие. Набор данных UNSW_NB15 содержит различные атрибуты, описывающие сетевой трафик, и метки классов, указывающие на тип атаки или отсутствие атаки для каждого экземпляра. Эти данные часто используются для обучения моделей машинного обучения с целью обнаружения и классификации сетевых атак.

Для реализации данного исследования был выбран язык программирования Python и библиотека scikit-learn для реализации Isolation Forest, а также Pandas для работы с данными. В рамках исследования выполнен анализ метода обнаружения аномалий с применением Isolation Forest, который является одним из методов несупервизионного обучения и применяется для выявления аномальных или необычных паттернов в данных, не требуя подробной разметки. Основные этапы проведенного исследования:

1. Загрузка данных и предобработка: данные загружаются из файлов и подвергаются необходимой предобработке, включая обработку категориальных признаков и масштабирование числовых.

2. Обучение модели Isolation Forest: создается и обучается модель Isolation Forest на подготовленных данных для выявления аномалий.

3. Оценка производительности модели: модель оценивается на тестовом наборе данных с использованием метрик: точность, матрица ошибок и отчет о классификации.

Используемые технологии:

```
# Импорт необходимых библиотек и модулей
from sklearn.ensemble import IsolationForest  # Импорт Isolation Forest для обнаружения аномалий
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report  # Импорт метрик для оценки модели
from sklearn.preprocessing import MinMaxScaler  # Импорт MinMaxScaler для масштабирования признаков
```

```
import pandas as pd # Импорт библиотеки для работы с данными
# Загрузка данных
features = pd.read_csv('DataSet/NUSW-NB15_features.csv', encoding='latin-1')  # Загрузка файла с признаками
training_set = pd.read_csv('DataSet/UNSW_NB15_training-set.csv')  # Загрузка тренировочного набора данных
testing_set = pd.read_csv('DataSet/UNSW_NB15_testing-set.csv')  # Загрузка тестового набора данных
# Обработка категориальных признаков
categorical_columns = ['proto', 'service', 'state', 'attack_cat’] # Выбор категориальных признаков
training_set_encoded = pd.get_dummies(training_set, columns=categorical_columns, drop_first=True)  # Кодирование категориальных признаков в тренировочном наборе
testing_set_encoded = pd.get_dummies(testing_set, columns=categorical_columns, drop_first=True)  # Кодирование категориальных признаков в тестовом наборе
common_columns = training_set_encoded.columns.intersection(testing_set_encoded.columns)  # Нахождение общих столбцов в обоих наборах данных
training_set_encoded = training_set_encoded[common_columns] # Оставление только общих столбцов в тренировочном наборе
testing_set_encoded = testing_set_encoded[common_columns] # Оставление только общих столбцов в тестовом наборе
# Масштабирование числовых признаков
scaler = MinMaxScaler()  # Инициализация MinMaxScaler
numerical_columns = training_set.select_dtypes(include=['float64', 'int64']).columns.tolist()  # Выбор числовых признаков
training_set_encoded[numerical_columns] = scaler.fit_transform(training_set[numerical_columns]) # Масштабирование числовых признаков в тренировочном наборе
testing_set_encoded[numerical_columns] = scaler.transform(testing_set[numerical_columns])  # Масштабирование числовых признаков в тестовом наборе
# Выбор признаков и целевой переменной
X_train = training_set_encoded.drop('label', axis=1)  # Отделение признаков от меток классов в тренировочном наборе
y_train = training_set_encoded['label’] # Выделение меток классов в тренировочном наборе
# Инициализация и обучение модели Isolation Forest
clf = IsolationForest(random_state=42) # Инициализация модели Isolation Forest с заданным random_state для воспроизводимости
clf.fit(X_train) # Обучение модели на тренировочных данных
# Прогнозирование аномалий на тестовом наборе
X_test = testing_set_encoded.drop('label', axis=1)  # Отделение признаков от меток классов в тестовом наборе
y_pred = clf.predict(X_test)  # Предсказание аномалий на тестовых данных с помощью обученной модели
# Преобразование предсказанных значений
y_pred[y_pred == 1] = 0  # Замена значений для представления аномалий (1)
y_pred[y_pred == -1] = 1  # Замена значений для представления нормальных данных (0)
# Оценка производительности модели Isolation Forest
accuracy = accuracy_score(testing_set_encoded['label'], y_pred) # Вычисление точности модели
conf_matrix = confusion_matrix(testing_set_encoded['label'], y_pred)  # Вычисление матрицы ошибок
classification_rep = classification_report(testing_set_encoded['label'], y_pred)  # Создание отчета о классификации
# Вывод результатов
print(f"Isolation Forest - Accuracy: {accuracy}")  # Вывод точности модели
print(f"Isolation Forest - Confusion Matrix:\n{conf_matrix}")  # Вывод матрицы ошибок
print(f"Isolation Forest - Classification Report:\n{classification_rep}")  # Вывод отчета о классификации
```

```
        # Обучение модели Isolation Forest
clf.fit(X_train)
# Прогнозирование аномалий на обучающем наборе для визуализации процесса обучения
y_train_pred = clf.predict(X_train)
# Создание графика для отображения прогнозирования аномалий на обучающем наборе
plt.figure(figsize=(10, 6))
plt.scatter(X_train.iloc[:, 0], X_train.iloc[:, 1], c=y_train_pred, cmap='viridis', label='Anomaly (-1) / Normal (1)')
plt.title('Isolation Forest - Training Set Anomaly Detection')
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.legend()
plt.colorbar()
plt.show(
```

Обсуждение результатов. Результат работы программного кода приведен на рис. 1.

Эти метрики и отчет о классификации показывают оценку производительности модели Isolation Forest на тестовом наборе данных для задачи обнаружения аномалий.

Accuracy (Точность): 0.32

Это общая точность модели, т.е., доля правильно классифицированных случаев.

Здесь она составляет около 32%, что может быть низким значением, особенно для некоторых задач классификации. Однако, для задачи обнаружения аномалий высокая точность может быть менее важной из-за дисбаланса классов.

Confusion Matrix (Матрица ошибок): в матрице ошибок видно, что модель правильно классифицировала 55174 аномальных случая и 833 нормальных случая.

Classification Report (Отчет о классификации):

Precision (Точность) для класса 0.0 (аномалии) составляет 0.32, что означает, что из всех предсказанных моделью аномалий только 32% действительно являются аномалиями.

Recall (Полнота) для класса 0.0 составляет 0.99, что означает, что модель обнаруживает 99% всех действительных аномалий.

F1-score (F-мера) для класса 0.0 равен 0.48, это среднее гармоническое между точностью и полнотой.

Support (Поддержка):

Это количество фактических случаев в каждом классе. Эти метрики помогают понять, как модель работает в контексте обнаружения аномалий. В данном случае, хотя точность невысока, полнота высока для аномалий, что может быть важным в контексте обнаружения редких событий, но стоит обращать внимание на ошибки в классификации нормальных случаев как аномалий.

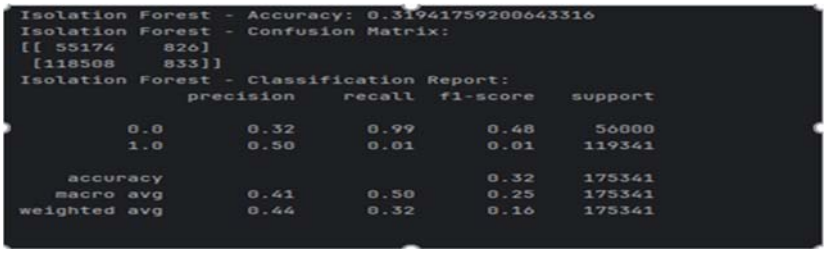

Рис. 1. Оценка производительности Isolation Forest
Fig. 1. Isolation Forest performance evaluation

График, созданный в данном коде, позволяет визуализировать процесс обучения модели Isolation Forest на обучающем наборе данных (рис. 2). Каждая точка на графике представляет отдельный образец из обучающего набора данных, где оси X и Y представляют различные признаки.

Синий (значение 1): Эти точки представляют образцы, которые модель считает «нормальными», то есть они не являются аномальными по мнению модели.

Желтый (значение -1): Точки этого цвета представляют образцы, которые модель отметила, как «аномалии» или необычные паттерны данных.

Таким образом, график позволит визуально оценить, как модель разделяет данные на «нормальные» и «аномальные» значения на обучающем наборе данных. Это позволяет оценить способность модели выявлять аномалии и ее общую производительность в их обнаружении.

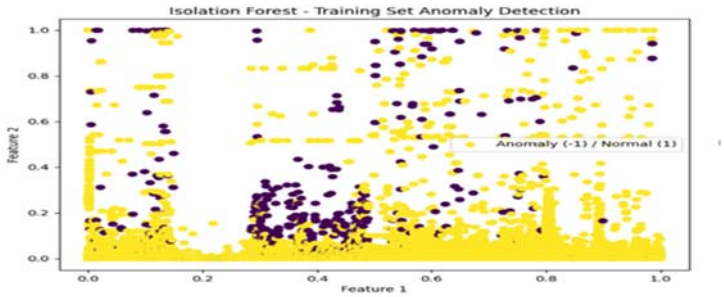

Рис. 2. Визуализации прогнозирования аномалий
Fig. 2. Visualizations of anomaly prediction

Вывод. Проделанная работа позволила исследовать метод Isolation Forest в контексте обнаружения аномалий в данных без детальной разметки. Оценка применимости данного метода на неструктурированных данных выявила его потенциал в выделении аномальных паттернов без необходимости обширной разметки. Это подтверждает эффективность Isolation Forest в условиях, где доступ к размеченным данным ограничен или отсутствует. Isolation Forest имеет достаточно высокую производительность при обнаружении аномалий в различных типах данных, включая неструктурированные данные, что делает его перспективным методом в областях, где требуется оперативное выявление нештатных ситуаций.

Оценка производительности модели Isolation Forest через различные метрики, такие как точность, полнота и F-мера, подтвердила его способность точно определять аномалии. Метод проявил свою применимость в реальных сценариях, таких как кибербезопасность или финансовая аналитика, что подчеркивает его значимость в областях, где оперативное обнаружение аномалий имеет высокий приоритет.

Таким образом, исследование показало, что Isolation Forest представляет собой перспективный метод для обнаружения аномалий в данных без обширной разметки и может успешно применяться в различных областях, где обнаружение нештатных ситуаций играет ключевую роль. Выводы работы могут быть использованы для дальнейшего улучшения методов обнаружения аномалий и их применения в реальных сценариях.

Библиографический список:

1. Попова, И.А. Обнаружение аномалий в наборе данных с помощью алгоритмов машинного обучения без учителя Isolation Forest и Local Outlier Factor / И.А. Попова // StudNet. – 2020. – Т. 3, № 12. – С. 1460-1470. – EDN XILRBX.
2. Гайдук, К.А. К вопросу о реализации алгоритмов выявления внутренних угроз с применением машинного обучения / К.А. Гайдук, А.Ю. Исхаков // Вестник СибГУТИ. – 2022. – Т. 16, № 4. – С. 80-95. – DOI 10.55648/1998-6920-2022-16-4-80-95. – EDN SGBSIH.
3. Савицкий, Д.Е. Выявление аномалий при обработке потоковых данных в реальном времени / Д.Е. Савицкий, М.Е. Дунаев, К.С. Зайцев // International Journal of Open Information Technologies. – 2022. – Т. 10, № 6. – С. 70-76. – EDN IGAWAO.
4. Терских, М. Г. Обнаружение аномального поведения пользователей в журналах событий безопасности Windows с применением алгоритмов машинного обучения / М. Г. Терских, Е. М. Тишина // Теория и практика современной науки. – 2018. – № 5(35). – С. 821-839. – EDN UYMTHC.
5. Динамическая аутентификация пользователей на основе анализа работы с компьютерной мышью / А. В. Березникер, М. А. Казачук, И. В. Машечкин [и др.] // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. – 2021. – № 4. – С. 3-16. – EDN XIQNIZ.
6. Токарев, Д. М. Обнаружение аномалий на основе машинного обучения с использованием сочетания алгоритмов K-MEAN и SMO / Д. М. Токарев, М. Г. Городничев // Телекоммуникации и информационные технологии. – 2023. – Т. 10, № 1. – С. 5-13. – EDN ILCJZP.
7. Мельник, М. В. Обнаружение аномального поведения пользователей и сущностей в контейнерных системах на основе методов машинного обучения / М. В. Мельник, И. В. Котенко // Информационная

безопасность регионов России (ИБРР-2023) : XIII Санкт-Петербургская межрегиональная конференция. Материалы конференции, Санкт-Петербург, 25–27 октября 2023 года. – Санкт-Петербург: Санкт-Петербургское Общество информатики, вычислительной техники, систем связи и управления, 2023. – С. 97-98. – EDN QOBTZP.

8. Н. Эйб, Б. Задрозный, Дж. Лэнгфорд. Обнаружение выбросов с помощью активного обучения. В материалах 12-й международной конференции ACM SIGKDD по обнаружению знаний и интеллектуальному анализу данных, страницы 504-509. ACM Press, 2006.
9. Сафин, А. Р. Обнаружение аномального поведения сетевого трафика на основе статистических методов при помощи машинного обучения / А. Р. Сафин // Информационная безопасность и защита персональных данных. Проблемы и пути их решения: Материалы XIII Межрегиональной научно-практической конференции, Брянск, 30 апреля 2021 года. – Брянск: Брянский государственный технический университет, 2021. – С. 228-231. – EDN UDRGDA.
10. А. Асунсьон, Д. Ньюман. Репозиторий машинного обучения UCI, 2007.

References

1. Popova, I.A. Detection of anomalies in a data set using unsupervised machine learning algorithms Isolation Forest and Local Outlier Factor/ I.A. Popova StudNet. 2020; 3(12):1460-1470. – EDN XILRBX. (In Russ)
2. Gaiduk, K.A. On the issue of implementing algorithms for identifying internal threats using machine learning / K.A. Gaiduk, A.Yu. Iskhakov. *Bulletin of SibGUTI.* 2022;16(4):P. 80-95. – DOI 10.55648/1998-6920-2022-16-4-80-95. – EDN SGBSIH. (In Russ)
3. Savitsky, D.E. Detecting anomalies when processing streaming data in real time / D.E. Savitsky, M.E. Dunaev, K.S. Zaitsev. *International Journal of Open Information Technologies*. 2022;10(6):70-76. – EDN IGAWAO. (In Russ)
4. Terskikh, M. G. Detection of anomalous user behavior in Windows security event logs using machine learning algorithms / M. G. Terskikh, E. M. Tishina. *Theory and practice of modern science*. 2018; 5(35): 821-839. – EDN UYMTHC. (In Russ)
5. Dynamic user authentication based on analysis of work with a computer mouse / A. V. Berezniker, M. A. Kazachuk, I. V. Mashechkin [etc.]. *Bulletin of Moscow University. Episode 15: Computational mathematics and cybernetics*. 2021; 4: 3-16. – EDN XIQNIZ. (In Russ)
6. Tokarev, D. M. Anomaly detection based on machine learning using a combination of K-MEAN and SMO algorithms / D. M. Tokarev, M. G. Gorodnichev. *Telecommunications and information technologies.* 2023; 10(1):5-13. – EDN ILCJZP. (In Russ)
7. Melnik, M. V. Detection of anomalous behavior of users and entities in container systems based on machine learning methods / M. V. Melnik, I. V. Kotenko. *Information security of regions of Russia (IBRR-2023): XIII St. -Petersburg interregional conference. Conference materials*, St. Petersburg, October 25–27, 2023. – St. Petersburg: St. Petersburg Society of Informatics, Computer Science, Communication and Control Systems, 2023: 97-98. – EDN QOBTZP. (In Russ)
8. N. Abe, B. Zadrozny, J. Langford. Outlier detection using active learning. In Proceedings of the 12th ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*, ACM Press, 2006: 504–509. (In Russ)
9. Safin, A. R. Detection of anomalous behavior of network traffic based on statistical methods using machine learning. Information security and personal data protection. Problems and ways to solve them: Materials of the XIII Interregional Scientific and Practical Conference, Bryansk, April 30, 2021. – Bryansk: Bryansk State Technical University, 2021: 228-231. – EDN UDRGDA. (In Russ)
10. Asuncion, D. Newman. UCI Machine Learning Repository, 2007. (In Russ)

Сведения об авторах:

Кечеджиев Александр Сергеевич, магистрант кафедры «Вычислительные системы и информационная безопасность»; Kechedzhiev.alex@mail.ru

Цветкова Ольга Леонидовна, кандидат технических наук, доцент, доцент кафедры «Вычислительные системы и информационная безопасность»; olga_cvetkova@mail.ru

Information about authors:

Alexander S. Kechedzhiev, Master's Student, Department of Computer Systems and Information Security; Kechedzhiev.alex@mail.ru

Olga L. Tsvetkova, Cand. Sci. (Eng), Assoc. Prof., Assoc. Prof., Department of Computer Systems and Information Security; olga_cvetkova@mail.ru

Конфликт интересов/Conflict of interest.

Авторы заявляют об отсутствии конфликта интересов/The authors declare no conflict of interest.

Поступила в редакцию/ Received 16.12.2023.

Одобрена после рецензирования / Reviced 17.01.2024.

Принята в печать /Accepted for publication 17.01.2024.