УДК 004.056
DOI: 10.55648/1998-6920-2022-16-4-80-95

К вопросу о реализации алгоритмов выявления внутренних угроз с применением машинного обучения[1]

К. А. Гайдук, А. Ю. Исхаков

В работе представлен анализ алгоритмов и подходов, применяемых при решении задачи для выявления внутренних угроз с применением методов машинного обучения. Выявление внутренних угроз в контексте данного исследования сводится к решению задачи детектирования аномалий в журналах аудита действий субъектов доступа. В статье формализованы основные направления выявления внутренних угроз и приведены популярные алгоритмы машинного обучения. В работе поднимается проблема объективной оценки результатов исследований и разработок в данной предметной области. На основании проведенного анализа разработаны рекомендации по реализации систем выявления внутренних угроз с помощью алгоритмов машинного обучения.

Ключевые слова: внутренние угрозы информационной безопасности, машинное обучение, поиск аномалий, аутентификация, изоляционный лес, ансамблевые методы.

1. Введение

Внутренние угрозы можно определить как любые потенциально опасные для организации действия субъекта (инсайдера). К их числу относятся несанкционированная передача данных, нарушение целостности ресурсов и другие злонамеренные или непреднамеренные действия. Задача выявления таких угроз усложняется тем, что инсайдеры способны тщательно скрывать следы и пытаться моделировать нормальное поведение. В связи с этим для эффективного решения задачи необходимо проанализировать большой объем данных системных журналов, зачастую в режиме реального времени. Внутренние угрозы принято разделять на две категории [1]: умышленные (злонамеренные) и непреднамеренные (пассивные). Пассивная внутренняя угроза происходит, когда сотрудник неосознанно предоставляет доступ злоумышленнику, игнорируя политику безопасности или пользуясь ее недостатками, посещая фишинговые сайты, переправляя конфиденциальные данные на личную электронную почту и т.д. Злонамеренные угрозы исходят от лиц, которые осознанно используют свой доступ к ресурсам компании, чтобы модифицировать, уничтожать или собирать конфиденциальную информацию в личных целях.

Подходы к обнаружению внутренних угроз можно разделить на три направления: на основе правил, на основе графов и на основе методов машинного обучения [2]. Подход, основанный на применении методов машинного обучения, получил широкое применение благодаря своему мощному математическому аппарату и высокой эффективности [3]. В контексте

[1] Исследование выполнено при частичной финансовой поддержке РНФ в рамках научного проекта № 21-71-00125 «Алгоритмическое обеспечение для усиленной проверки подлинности субъектов доступа в критически важных объектах».

внутренних угроз машинное обучение используется для создания моделей, идентифицирующих угрозы и статистически определяющих подозрительное поведение. Классификация методов решения задачи выявления внутренних угроз не может быть однозначной, так как количество этих методов постоянно растет и повторно анализируется в совокупности. Более подробная классификация приведена в [4].

Целью данной работы является исследование тенденций применения методов машинного обучения в реализации алгоритмического обеспечения для выявления внутренних угроз. Для этого предлагается: провести систематизацию основных подходов; провести обзор современных исследований в этой области; выделить проблемные области, с которыми сталкиваются разработчики подобных систем.

2. Подходы к построению систем выявления внутренних угроз

В контексте машинного обучения задача выявления внутренних угроз формулируется как задача обнаружения аномалий, которые в зависимости от методов способны выявить как атаки по известному сценарию, так и новые типы атак [5]. Выявление аномалий – это обнаружение выбросов, редких событий и наиболее отличающихся значений в наборе данных. Задача выявления атак по известному сценарию чаще решается эффективнее и быстрее, но вместе с этим эти сценарии необходимо непрерывно обновлять, что создает зависимость от экспертной оценки.

Система выявления аномалий должна обрабатывать большие объемы информации, зачастую это системные и сетевые журналы, отображающие действия пользователей. Методы машинного обучения получают на вход преобразованные в вектор признаков данные и обучаются на них, чтобы затем принять решение, является ли такой вектор признаков аномальным. Таким образом, задача выявления аномалий ставит ряд вопросов о выборе парадигмы обучения, конкретного метода и вопрос представления данных.

Первый вопрос для специалистов – это выбор формата данных, в которых будет осуществляться поиск внутренних угроз. Понимание специфики работы отслеживаемой системы необходимо для выбора достоверного источник данных, который зачастую представляет собой журналы активности пользователей, собранные из различных источников.

Предварительная обработка данных также является неотъемлемой частью анализа данных, целью которого является удаление нежелательного шума в данных и, таким образом, уточнение интересующих характеристик данных. Более того, выбор оптимальных методов предварительной обработки может существенным образом повлиять на результаты анализа.

После того как данные поступают от источников в различной форме, их необходимо обработать и привести к единой форме. Во-первых, данные совмещаются из различных источников на основе идентификатора пользователя с учетом условия этого совмещения, например, какой-то период времени или количество действий. Так как данные поступают от разных источников с разным числом полей, их нужно привести к одному виду с помощью выделения наиболее релевантных.

Поскольку алгоритмы машинного обучения работают с числовыми данными, необходимо кодирование признаков для создания числовых векторов. Ошибки в процессе кодирования переменных функций могут привести к тому, что модели машинного обучения неправильно интерпретируют корреляцию между ними. Например, при кодировании ограниченного числа категорий у признаков целыми числами алгоритмами машинного обучения это может восприниматься как отношение порядка, тогда как в реальности это не так, поэтому признаки разделяют на категориальные и порядковые.

После приведения данных к единому виду необходимо провести нормализацию данных, к задачам которой можно отнести сокращение объема данных, удаление повторных записей, противоречивых или некорректных данных. К алгоритму нормализации можно отнести метод синтетической переборки меньшинства (SMOTE), который позволяет сохранить

производительность при несбалансированном наборе классов в задаче классификации. Алгоритм заключается в выборе записи из класса меньшинства в качестве входных данных, выполняется поиск его k ближайших соседей и между этими точками генерируется одна или несколько новых записей.

После всех этапов обработки данные делятся на обучающую выборку, на которой алгоритм учится, и тестовую выборку для проверки точности модели.

Так как задача выявления аномалий относится к задачам классификации, то можно применить те же метрики для оценки качества модели, такие как точность (precision), полнота (recall), F-мера и анализ ROC-AUC-кривой.

Таким образом, обобщение процесса построения систем выявления внутренних угроз, основанных на методах машинного обучения, можно представить в виде концептуальной схемы на рис. 1.

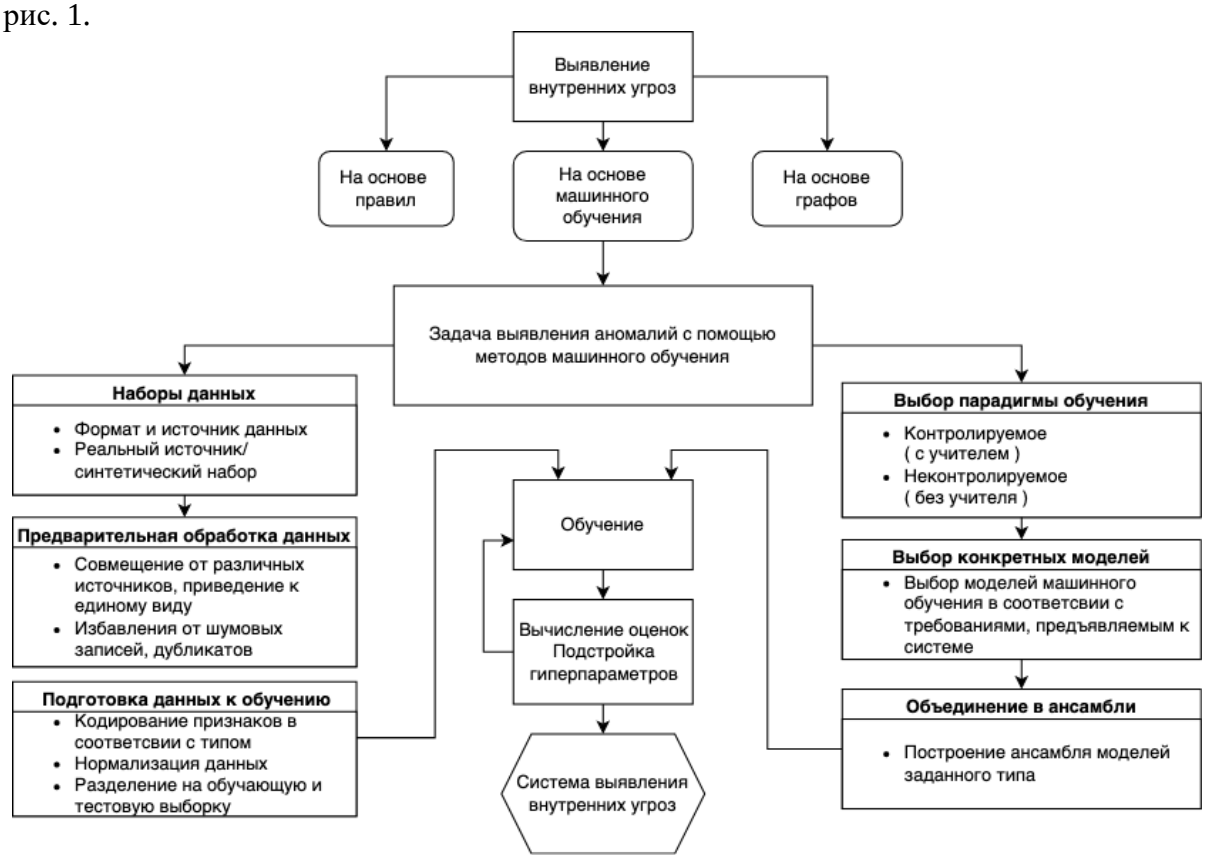

Рис.1. Схема построения системы выявления внутренних угроз

3. Методы машинного обучения, применяемые для решения задачи выявления внутренних угроз

3.1. Контролируемые и неконтролируемые методы поиска аномалий

В машинном обучении принято выделять два направления поиска аномалий: обнаружение выбросов и обнаружение новизны. Выбросы и новое поведение легко спутать, так как они являются объектами, отличающимися по своим свойствами от объектов обучающей выборки, именно поэтому важно правильно формулировать задачи для выбора методов и формировать обучающую выборку с учетом поставленных задач. В контексте выявления внутренних угроз стоит задача выявления выбросов, поэтому данные, на которых будет обучаться модель, должны обладать достаточной полнотой и описывать максимально большое множество «нормальных» поведений в системе.

Задача выявления атак по известному сценарию чаще решается эффективнее и быстрее, но вместе с этим сценарии необходимо непрерывно обновлять, что будет создавать зависимость от экспертной оценки.

В машинном обучении существует две парадигмы: контролируемые (с учителем) и неконтролируемые методы (без учителя), которые отличаются способом представления обучающих данных.

В контролируемых методах обучение происходит с заранее помеченными целевыми данными, когда заранее известно, являются ли действия вредоносными или не являются, тем самым увеличивая точность обнаружения.

Одним из наиболее популярных контролируемых методов является метод k ближайших соседей (KNN), который используется для задач классификации [6].

Этот метод настраивает информацию о соседних точках для классификации выходных меток. Данный метод широко используется в области безопасности, например, для обнаружения вторжений и обнаружения спама и вредоносных программ [7]. Как правило, алгоритм KNN выбирает ближайшие выборки на основе измерения расстояния между всеми выборками обучающей выборки и новыми выборками. Он имеет возможность прогнозировать целевой класс более простым способом и с высокой точностью и демонстрирует значительную производительность при зашумленных обучающих данных или массивном наборе данных.

Неконтролируемые методы имеют дело с неразмеченными данными, обучение которых обычно требует более длительного времени и меньшей точности за счет выявления любых аномальных действий, не обязательно несущих угрозу. Наибольшее применение из неконтролируемых методов получили одноклассовый метод опорных векторов (OSVM) и изоляционный лес (IF).

Метод опорных векторов (SVM) изначально был разработан как алгоритм классификации двух или более классов, который находит максимальную границу, разделяющую классы, и является контролируемым алгоритмом [8]. Разработанный алгоритм был способен работать только с линейно разделимыми классами, однако получил большую популярность с введением «Kerneltrick», который позволил эффективно работать с линейно неразделимыми данными. Такой подход потенциально позволяет получать бесконечноразмерные пространства признаков. «Kerneltrick» основан на идее, что линейно неразделимые признаки в заданном пространстве могут стать разделимыми в пространствах большей размерности.

Одноклассовый SVM предназначен для классификации выбросов в том случае, когда данные содержат только положительные точки [9], и может иметь лишь один класс. Общая идея состоит в преобразовании пространства признаков и разделении гиперплоскостью так, чтобы наблюдения были наиболее удалены от начала координат. Полученная граница разделяет близко расположенные наблюдения из выборки и аномальные значения. Таким образом, данный алгоритм обучается на «нормальных» данных и способен обучаться на неразмеченных данных.

Алгоритмы «деревья решений» являются наиболее часто используемыми методами машинного обучения. Решения раскладываются в древовидные структуры до тех пор, пока не будет принято решение об успешном предсказании. Данный класс алгоритмов получил свою популярность из-за достаточно большой скорости и точности при решении задач машинного обучения [10].

Одним из известных представителей деревьев решений является алгоритм «случайного леса», который заключается в использовании сразу нескольких решающих деревьев. В таком случае решение принимается на основе выходных данных от каждого дерева перед получением единственного ответа. Исходная выборка случайным образом делится на N подмножеств, на каждом подмножестве строится дерево решений для проверки соответствия одного или нескольких признаков. Для каждого ветвления в дереве выбирается несколько случайных признаков (для каждого ветвления свои признаки). Далее выбирается лучший признак и разветвление по нему, а дерево строится до исчерпания выборки – как правило, пока в листьях не останутся представители только одного класса.

Применительно к задаче выявления внутренних угроз используется алгоритм «Изоляционный лес» (Isolationforest, IF) (рис. 2) [11]. Основываясь на принципе, что примеры аномалий редки и значительно отличаются по значениям атрибутов от нормальных точек данных, IF разработан как ансамбль «деревьев изоляции», при этом предполагается, что аномалии, которые легче изолировать, находятся ближе к корням деревьев, чем нормальные экземпляры.

Этот алгоритм отличается от других методов обнаружения аномалий, которые строят модели (в основном) нормальных данных и определяют аномалии как любые экземпляры, не соответствующие модели. Каждое дерево в IF работает на подмножестве обучающих данных и наборе признаков. В каждом узле дерева генерируются двоичные разбиения по случайно выбранному признаку и значению разбиения. Процесс рекурсивно повторяется до тех пор, пока каждый экземпляр не будет изолирован от других. После обучения всех деревьев изоляции балл аномалии экземпляра данных рассчитывается как средняя длина пути от корневых узлов до соответствующих листьев экземпляра в деревьях [12].

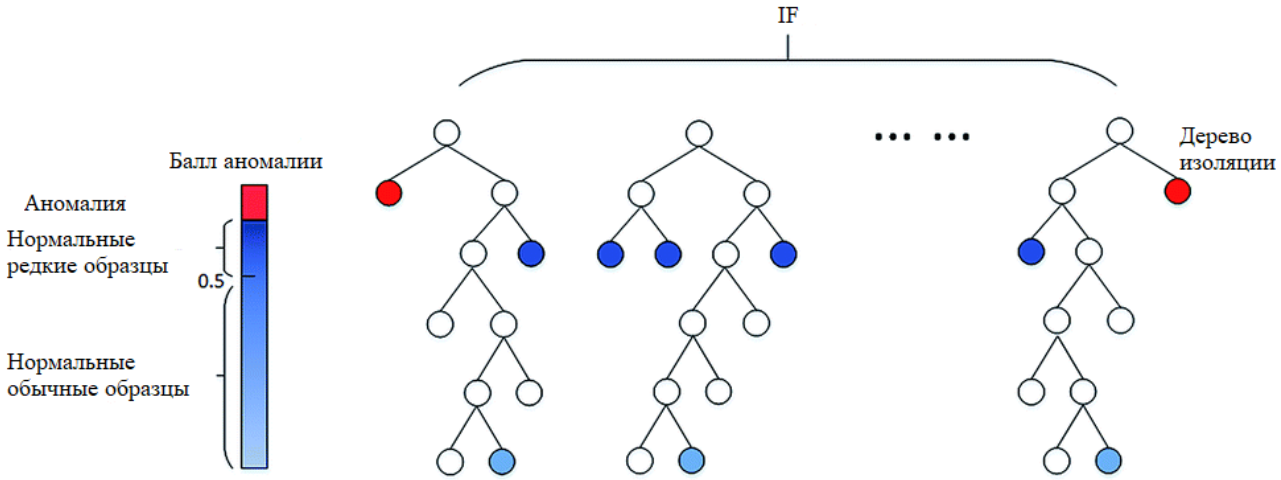

Рис. 2. Схема изоляционного леса

При правильном подборе количества деревьев и достаточно большой выборке данный алгоритм способен выявлять аномалии намного быстрее. Изначально разделяющие линии могут быть горизонтальными и вертикальными, что может существенно хуже работать на некоторых наборах данных. Для решения данной проблемы применяют модификацию расширенного изоляционного леса, который заключается в случайном выборе наклона ветви и точки пересечения [13].

В обоих случаях к обучающим данным предъявляется ряд требований для эффективной работы моделей, в то время как представление данных для модели может по-разному влиять на контролируемые и неконтролируемые модели.

Описанные выше модели машинного обучения являются не единственными, но наиболее популярными в контексте выявления аномалий. Неконтролируемые методы выявления аномалий работают в предположении, что большая часть данных нормальна, что позволяет в автоматическом режиме выявить нерегулярное и неестественное поведение пользователей. Например, одноклассовый SVM зачастую способен обучаться только на нормальных данных, что полезно в случае нехватки аномальных экземпляров данных, когда аномальное поведение в системе минимально. Изоляционный лес, напротив, способен выявлять аномалии с первых итераций своей работы и эффективен только при наличии аномалий в наборе данных, то есть, когда мы точно знаем, что в системе присутствуют аномалии.

Контролируемые методы требует заранее помеченных данных, что усложняет процесс предварительной обработки и накладывает ограничения на применение в некоторых системах, в то время как такие методы способны выявлять более «трудные» аномалии и обнаруживать даже тщательно скрытое аномальное поведение. Алгоритм *k* ближайших соседей способен эффективно работать в случае системы с большим количеством атрибутов с большой точность,

а также может применяться для обработки данных, дополняя или восстанавливая данные в случае использования метода для регрессии.

3.2. Ансамблевые методы

При выявлении угроз с помощью алгоритмов машинного обучения основной задачей исследователя является минимизация ошибок первого и второго рода. При классификации объекта как «нормальный» или «аномальный» с помощью его признаков и используя конкретную модель машинного обучения специалисты сталкиваются с типичными недостатками и ошибками этой модели. Объединив несколько таких моделей (обычно называемых «слабыми»), обучая их для решения одной и той же задачи, получают ансамбль методов.

Основная гипотеза ансамблевых методов состоит в том, что при определенном сочетании слабых моделей можно получить более точные, «сильные» модели. Под слабыми моделями понимают те, для которых точность может быть чуть выше случайного угадывания, в то время как сильные модели позволяют добиться в теории произвольно малой ошибки [14].

Существует множество ансамблевых методов как для определенных моделей, так и для общих принципов построения ансамблей. Наиболее популярными являются стекинг (суммарное обобщение), бэггинг (настройка с объединением) и бустинг (улучшение).

Бэггинг (от англ. Bootstrap Aggregation) обучает модели на разных подмножествах исходного обучающего множества независимо, в отличие от стэкинга (от англ. Stacked Generalization – стековое обобщение), который обучает принципиально различные модели на двух частях исходного обучающего множества: на первом модели обучаются, на втором проверяется их работа, затем выходы от моделей используются в качестве входных данных для мета-модели. Для задач классификации, к которым можно отнести выявления аномалий, наибольшую популярность получил бустинг.

Бустинг (от англ. boosting – улучшение) – это последовательный алгоритм композиции слабых моделей машинного обучения, реализующий сильную модель. Каждая последующая модель в композиции стримится компенсировать недостатки композиции предыдущих моделей. Важно отметить, что используемые модели в композиции должны быть действительно слабыми для улучшения общей модели в результате композиции.

Одним из первых алгоритмов бустинга является AdaBoost (сокращение от adaptive boosting), который также использует композицию слабых моделей. Обучающая выборка на каждой итерации композиции определяется исходя из ошибок на предыдущих итерациях, то есть ошибка используется как вес для обучающих данных таким образом, что большие веса присваиваются тем примерам, на которых ошибались предыдущие модели.

Для решения проблем адаптивного бустинга на основе этого алгоритма появилось обобщение – градиентный бустинг (GBM). Градиентный бустинг определяет слабые модели на основе градиентов функции потерь, которая является мерой отклонения выходных значений модели от реального значения. Таким образом, задача формулируется как обучение с учителем, а именно, необходимо выбрать функцию потерь и минимизировать ее с помощью градиентных методов. В разных задачах используются разные функции потерь, и от ее выбора зависит качество всей модели.

Одним из популярнейших на сегодняшний день алгоритмов градиентного бустинга является XGBoost (от англ. eXtreme Gradient Boosting), который работает с деревьями решений в качестве слабых моделей и представлен в виде программной библиотеки с открытым исходным кодом для большинства языков программирования. Так же, как и в общей идее градиентного бустинга, на каждой итерации вычисляется отклонение (значение целевой функции) уже обученного ансамбля на обучающей выборке. Следующая композиция добавляется в модель с целью предсказания этих отклонений и уменьшения среднего отклонения всей модели. Новые модели будут добавляться до тех пор, пока ошибка уменьшается (достаточный положительный градиент) либо пока не выполнится одно из правил досрочной остановки.

3.3. Наборы данных

В настоящее время не существует общедоступного набора реальных данных для обнаружения внутренних угроз, поэтому исследователи в своих работах применяют синтетически сгенерированные данные. Большинство наборов представляют собой данные в формате *csv*, строки которых представляют записи – обычно это дата, время, информация о пользователе или устройстве и непосредственно действие в системе.

Один из первых таких наборов был представлен в 1998 году лабораторией Линкольна Массачусетского технологического института для программы оценки обнаружения вторжений DARPA [15]. Этот набор представляет собой около пяти миллионов записей TCP-соединений за семь недель работы системы.

Набор данных RUU был собран Малеком Бен Салемом с 34 компьютеров под управлением операционной системой Windows. Этот набор включает в себя действия в системе, реестр Windows, доступ к файловой системе, запущенные процессы и используемые динамические библиотеки [16].

Набор данных TWOS [17] был собран из реальных взаимодействий пользователя с хост-машиной. Набор был собран во время конкурса, организованного Сингапурским университетом технологии и дизайна в марте 2017 года, и включает в себя данные, собранные из шести источников данных (нажатия клавиш, мышь, монитор хоста, сетевой трафик, журналы SMTP и вход в систему), а также дополнительные психометрические данные.

Одним из самых популярных наборов данных является CERT [18], который был создан Институтом программной инженерии Университета Карнеги-Меллона. Это «свободный от ограничений конфиденциальности» [19] набор, разработанный, чтобы позволить исследователям, изучающим тему внутренних угроз, экспериментировать и оценивать предлагаемые ими решения. Он имеет несколько различных версий, данные в каждой версии генерируются с использованием различных сценариев. Каждый набор данных содержит журналы данных входа в систему, HTTP или историю просмотров, электронную почту, журналы доступа к файлам, использование устройств, данные LDAP, а также дополнительную психометрическую информацию.

В рамках исследования был проведен обзор существующих наборов данных, в том числе частоты и особенности их применения исследователями. Проведенный анализ позволил выявить подходы к составлению наборов и основные признаки, что подтвердило неоднозначность решения проблемы выявления внутренних угроз.

Из проведенного анализа однозначно можно выделить проблемы отсутствия реальных наборов из-за вопросов конфиденциальности и отсутствия унификации среди различных наборов данных.

4. Результаты анализа современных исследований и рекомендации к системам выявления внутренних угроз

На сегодняшний день проводится целый ряд научных исследований, в которых предлагается различное методическое и алгоритмическое обеспечение для задачи обнаружения внутренних угроз в действиях пользователей. В большинстве публикаций предлагаются подходы к агрегации, корреляции и нормализации событий в журналах аудита, а также оптимизации известных алгоритмов машинного обучения с целью достижения лучших показателей работы систем обнаружения угроз. Оценку работы методов в известных публикациях стараются проводить на общеизвестных наборах данных. Одним из наиболее популярных и полным по своей структуре является набор CERT [18].

В связи с этим в данной работе с целью проведения объективного сравнения был также выбран набор данных CERT с заданным набором признаков. Проведенный анализ свидетельствует о том, что в зависимости от версии данного набора он содержит разное количество записей и внутренних угроз, но структура данного набора остается единой и содержит 5 разделенных таблиц: данные HTTP, EMAIL, данные входа и выхода в систему, действия с файлами, данные о подключениях устройств. В каждой из таблиц содержится информация о пользователе, идентификаторе действия, дате и времени и другие поля, специфичные для данной таблицы.

Ниже представлен анализ результатов, полученных в некоторых публикациях с набором данных CERT. Прежде всего, следует отметить результаты с разными методами кодирования и предварительной обработки. В [20] авторы сравнивают алгоритмы логистической регрессии (LR), деревьев решений (DT), случайного леса (RF), KernelSVM (KSVM) и *k* ближайших соседей (KNN) на наборе CERT r4.2. В работе применяются методы масштабирования переменных и метод синтетической переборки меньшинства. Используя эти методы в отдельности, авторы сравнивают полученные оценки: *F*-меру, точность и полноту (табл. 1, 2).

Таблица 1. Сравнение оценок для исходных и масштабированных данных [20]

Алгоритм	Точность		Полнота		*F*-мера	
	Исходные	Масштаб	Исходные	Масштаб	Исходные	Масштаб
DT	0.67	0.67	0.74	0.74	0.71	0.71
RF	0.60	0.58	0.19	0.19	0.29	0.29
KNN	0.13	0.84	0.01	0.21	0.02	0.32
KSVM	0.00	1.00	0.00	0.15	0.00	0.27

Из табл. 1 видно, что методы KSVM и KNN неустойчивы к ненормированным данным, поэтому их оценки увеличиваются при масштабировании данных, в то время как оценки других методов изменяются незначительно.

Таблица 2. Сравнение оценок для исходных и сбалансированных данных методом синтетической переборки меньшинства данных [20]

Алгоритм	Точность	Полнота	F-мера
Без использования метода SMOTE			
LR	1.00	0.15	0.26
DT	0.68	0.77	0.72
RF	0.39	0.32	0.35
KNN	0.33	0.06	0.11
С использованием SMOTE			
LR	0.50	1.00	0.67
DT	0.99	0.99	0.99
RF	0.99	0.99	0.99
NB	0.77	0.95	0.85
KNN	0.98	0.99	0.98

Для такого сравнения авторы использовали метод синтетической переборки меньшинства (SMOTE): генерировали новые записи и сравнили обученные модели на исходном и дополненном множествах данных, закодированных методом «горячего кодирования». Как видно из табл. 2, у большинства методов оценки увеличились, что говорит об их неустойчивости к несбалансированности в данном случае. Справедливо заметить, что данное улучшение в работе моделей возможно только с заранее размеченными данными для возможности применить SMOTE.

В качестве финального результата авторами предлагается рис. 3, где по оси ординат отмечены методы, по оси абсцисс – методы кодирования (метод синтетической переборки меньшинства, метод «горячего кодирования», кодирование меток), а по оси аппликат – значение ROC-AUC.

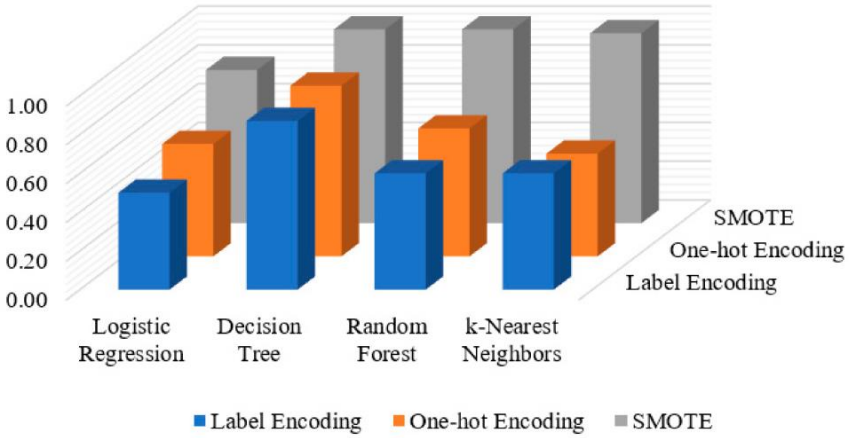

Рис. 3. Результат сравнительного анализа [20]

Таким образом, авторы называют методы деревьев решений и случайного леса наиболее надежными по сравнению с другими алгоритмами, что говорит об их преимуществе в условиях несбалансированности и зашумленности данных.

В работе [21] также используется синтетическая переборка меньшинства и сравниваются метод XGBoost и метод случайного леса на наборе CERT r6.2. Из анализа табл. 3 можно видеть очевидное преимущество метода XGBoost перед методом случайного леса, что также подтверждает тот факт, что ансамблевые методы имеют большую эффективность при оптимальной настройке гиперпараметров.

Таблица 3. Сравнение оценок для двух моделей

Метод	Accuracy(10)	Точность	Полнота	*F*-мера
RandomForest	77.10 %	86.82 %	90.41 %	88.58 %
XGBoost	99.13 %	98.34 %	98.21 %	98.27 %

Отмечается, что в публикациях часто сравнивают методы машинного обучения и искусственные нейронные сети. Например, в [22] сравниваются методы случайного леса (RF), которые используют алгоритм CART (деревья классификации и регрессии с учетом индекса «Джинни»), логистическая регрессия (LR) и искусственная нейронная (ANN) сеть прямого распространения с тремя скрытыми слоями, для которой применяется алгоритм оптимизации ADAM. Авторы используют наборы данных CERT r5.1 и CERT r6.2 и сравнивают алгоритмы на них. В работе также сравниваются два подхода к выявлению угроз – выявление угроз по конкретному экземпляру (записи) и выявление угрозы по пользователю. В качестве параметров оценки качества выбраны полнота и точность (табл. 4).

Таблица 4. Сравнение оценок для разных представлений данных

Алгоритм	Полнота (Recall)	Точность (Precision)
На основе экземпляров		
LR	0.5752	0.171
RF	0.4642	0.994
ANN	0.4676	0.391

На основе пользователей		
Алгоритм	Полнота (Recall)	Точность (Precision)
LR	0.9323	0.284
RF	0.6954	0.996
ANN	0.7815	0.532

Из табл. 4 видно, что подход выявления внутренних угроз на основе действий пользователя, а не всех действий в отдельности дает большое преимущество. В результате авторы выделяют RF и ANN как модели, показывающие лучшие результаты, и рекомендуют использовать RF в случае ограниченного набора данных с большей точностью, с другой стороны, использование ANN дает больший параметр полноты.

В табл. 5 представлен сравнительный обзор, систематизирующий результаты исследований, посвященных задаче выявления внутренних угроз. Были взяты максимальные оценки, полученные авторами с использованием наборов данных CERT разных версий. Стоит отметить, что во всех перечисленных работах размер обучающей выборки разнится, к тому же зачастую записи объединяются в одну за промежуток времени, что объясняет разность некоторых результатов.

Таблица 5. Сравнительный анализ современных исследований

Ссылка	Набор данных	Оценки для моделей			Описание
[22]	r4.2, r6.2	*AUC*	*Recall*	*Accuracy*	Используется сразу два набора данных, которые модифицировались и объединялись для большей производительности. В работе строятся ансамбли из моделей при помощи голосования большинства. В работе предлагаются и другие ансамбли с нечётным количеством модели, приведена оценки для лучшего ансамбля. Отмечается чувствительность SVM к классификаторам.
		One-class SVM			
		0.95	1	0.265	
		Изоляционный лес			
		0.596	0.025	0.958	
		HMM+LOF+SVM:			
		0.966	1	0.909	
[23]	r4.2	*AUC*	*PR*		Определяется показатель доверия для пользователя за период времени. Сравнивается эффективность моделей за разные периоды, и для последующих периодов учитывается показатель доверия за предыдущий период.
		One-class SVM			
		0.97	0.96		
		Изоляционный лес			
		0.91	0.97		
[24]	r4.1, r4.2	*Accuracy*	*Среднеквадратичная ошибка*		Применяется метод нормализации. Сравниваются модели, обученные несколько раз и один раз на всём наборе данных, в результате отличий не
		Случайный лес			
		0.8975	0.2604		
		Наивный байесовский классификатор			
		91.966	0.2576		

		Модель 1 ближайшего соседа				выявлено. Сравниваются средняя и среднеквадратичная ошибки методов.
		94.682		0.2115		
[10]	r4.2, r6.2, другие	*AUC*		*Время обучения, с*	*Время прогноза, мс*	Сравниваются разные временные представления данных и представления на основе пользователей и экземпляров. Сравнение моделей в условиях «отравления» данных, с разным числом пользователей и периодом. Комбинирование показателей обнаружения с помощью ансамблей.
		Нейронная сеть «Автокодировщик»				
		0.936		200	<1	
		LOF				
		0.768		472	13	
		Изоляционный лес				
		0.744		40	<1.25	
		LODA				
		0.863		20	<1	
[25]	r6.2	*PR*	*Recall*	*F-мера*	*AUC*	Предложена стратегия обработки данных для решения несбалансированности и сравнение результатов с исходным набором и синтетически дополненным с помощью алгоритма SMOTE. Приведены оптимальные параметры обучения.
		Случайный лес				
		0.9678	0.7852	0.8669	0.8581	
		Случайный лес + SMOTE				
		0.9784	0.8626	0.9168	0.8916	
		XGboost				
		0.9857	0.8292	0.9	0.9178	
		XGboost + SMOTE				
		0.9951	0.9816	0.9883	0.9687	
[26]	r5.2, r6.2	*FPR*	*TPR*	*Precision*	*F-мера*	В работе делается акцент на сравнение различных вариаций обучающего множества: представление данных по пользователям или по записям, разные периоды времени, идеальные и реальные условия (разбиение данных на обучающее и тестовое множества). Используется несколько наборов данных с различными сценариями для проверки обобщающей способности.
		Логистическая регрессия				
		0.0775	0.9154	0.2817	0.43	
		Многослойный персептрон				
		0.0282	0.77	0.49	0.59	
		Случайный лес				
		0.0043	0.81	0.88	0.8442	
		XGboost				
		0.001	0.82	0.75	0.7825	
[27]	r5.2	*TPR*	*PR*	*F-мера*	*Время обучения, с*	Использование разного шага по времени, трех разных подходов для нормализации данных, разные подходы к представлению данных.
		Логистическая регрессия				
		0.8815	0.361	0.51	60.86	
		Случайный лес				
		0.85	0.7956	0.8156	42.17	
		Многослойный персептрон				
		0.9507	0.2756	0.4252	378.58	

Обозначения: PR – Precision (точность), HMM – скрытые марковские модели, LOF – локальный коэффициент выбросов, LODA – легкий интерактивный детектор аномалий.

Анализируя результаты публикаций при оценке базовых алгоритмов следует учитывать отсутствие единого подхода к проведению исследований у авторов в части предварительной обработки данных и настройки гиперпараметров, в связи с чем результаты различных коллективов отличаются. Эффективность подавляющего большинства базовых методов зависит от обработки набора данных: масштабирования, совмещения, разбиения на разные временные отрезки, баланса между классами и т.д. Ансамблевые методы в меньшей мере зависят от такой обработки, и в различных условиях такие методы обладают большей эффективностью в отличие от применения моделей отдельно [28].

Из анализа публикаций невозможно определить явного лидера из-за разности в условиях проведения экспериментов и оценки, однако результаты свидетельствуют о том, что методы изоляционного леса, случайного леса и одноклассовый метод опорных векторов более эффективны для использования отдельно. XGBoost и ансамблевые модели в целом позволяют улучшить эффективность практически любой модели, но также требуют тщательного подбора параметров.

На основе анализа публикаций можно выделить некоторые рекомендации для решения задачи выявления внутренних угроз:

1. Представление и обработка данных главным образом влияет на производительность любой модели, поэтому в данном случае следует уделять особенное время данному классу методов. В вопросе выбора данных большее значение имеет актуальность данных, включая систему учета прошлого опыта с сохранением периодичности и масштаба.
2. Система на основе моделей машинного обучения должна постоянно обучаться на основе журналов системы и предоставлять наглядные данные с сохранением корреляций для упрощения предупреждения и расследования инцидентов безопасности.
3. Выбор оценок для системы выявления внутренних угроз должен зависеть от поставленных задач и критичности инфраструктуры – это можно отнести к четкому пониманию архитектуры такой системы и привлечению специалистов.
4. Неконтролируемые модели машинного обучения помогают избавиться от высокой стоимости обслуживания такой системы, в то время как контролируемые методы обладают большей точностью и надежностью и оправдывают свои затраты. Невозможно добиться полной независимости от экспертной оценки на всем этапе функционирования системы выявления внутренних угроз, поэтому целесообразно использовать оба подхода к обучению моделей.
5. Надежная и эффективная система выявления внутренних угроз должна использовать комбинацию нескольких независимых моделей, объединяя их с помощью ансамблевых методов, которые, как показано на практике, превосходят любую отдельно взятую модель.

5. Заключение

Выявление внутренних угроз остается одной из самых сложных задач обеспечения безопасности для любой организации. В данной области значительные результаты принадлежат исследованиям применимости моделей на основе машинного обучения. Рассмотренные работы с помощью новейших подходов решают данную проблему эффективнее, чем классические подходы, что подтверждается практическими результатами.

При этом отсутствие единых критериев оценки и проблемы с воспроизводимостью результатов, которые вызваны недоступностью широкому кругу исследователей унифицированных наборов данных, порождает дополнительные вопросы и вызовы для современных ученых. В данной работе была сформулирована задача выявления внутренних угроз, также приведен анализ наиболее популярных контролируемых и неконтролируемых моделей на основе машинного обучения, описаны наборы данных и методы их обработки, необходимые для успешного решения поставленной задачи. Представлена систематизация существующих научных результатов и приведены рекомендации по реализации систем выявления внутренних угроз.

Описание, обзор, анализ и сравнение результатов по современным публикациям вместе с приведенным рекомендациями дает понимание современного состояния данной области исследований и задает направление для практических реализаций.

Исходя из анализа, можно выделить основные тенденции современных разработок, подтверждающие, что наиболее перспективным направлением в области выявления внутренних угроз являются искусственные нейронные сети, которые все чаще применяются в контексте решения рассматриваемых задач. Вместе с рассмотрением архитектур нейронных сетей необходимо учитывать специфику их работы, вопросы производительности и возможность реального применения в системах выявления внутренних угроз.Анализу этих и других вопросов, связанных с применением искусственных нейронных сетей в задаче выявления внутренних угроз, планируется посвятить дальнейшие исследования.

Литература

1. *Kim A., Oh J., Ryu J., Lee K.* A Review of Insider Threat Detection Approaches with IoT Perspective // IEEE Access. 2020. V. 8. P. 78847-78867.
2. *Kim J., Park M., Kim H., Cho S., Kang P.* Insider Threat Detection Based on user Behavior Modeling and Anomaly Detection Algorithms // Appl. Sci. 2019. V. 9, 4018.
3. *Alpaydin E.* Introduction to Machine Learning. MIT Press: Cambridge, MA, 2014.
4. *Al-Mhiqani M. N. et al.* A review of insider threat detection: Classification, machine learning techniques, datasets, open challenges, and recommendations // Applied Sciences. 2020. V. 10, № 15. P. 5208.
5. *Al-Mhiqani M. N. et al.* A new intelligent multilayer framework for insider threat detection // Computers & Electrical Engineering. 2022. V. 97. P. 107597.
6. *Rajaguru H., Chakravarthy S. R. S.* Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer // Asian Pacific journal of cancer prevention: APJCP. 2019. V. 20, № 12. P. 3777.
7. *Sarma M. S. et al.* Insider threat detection with face recognition and KNN user classification // IEEE International Conference on Cloud Computing in Emerging Markets (CCEM). 2017. P. 39–44.
8. *Chauhan V. K., Dahiya K., Sharma A.* Problem formulations and solvers in linear SVM: a review //Artificial Intelligence Review. 2019. V. 52, № 2. P. 803–855.
9. *Khan S. S., Madden M. G.* One-class classification: taxonomy of study and review of techniques // The Knowledge Engineering Review. 2014. V. 29, № 3. P. 345–374.
10. *Buczak A. L., Guven E.* A survey of data mining and machine learning methods for cyber security intrusion detection // IEEE Communications surveys & tutorials. 2015. V. 18, № 2. P. 1153–1176.
11. *Le D. C., Zincir-Heywood N.* Anomaly detection for insider threats using unsupervised ensembles // IEEE Transactions on Network and Service Management. 2021. V. 18, № 2. P. 1152–1164.
12. *Sadaf K., Sultana J.* Intrusion detection based on autoencoder and isolation forest in fog computing // IEEE Access. 2020. V. 8. P. 167059–167068.
13. *Hariri S., Kind M. C., Brunner R. J.* Extended isolation forest // IEEE Transactions on Knowledge and Data Engineering. 2019. V. 33, № 4. P. 1479–1489.
14. *Zhang C., Ma Y. (ed.)*. Ensemble machine learning: methods and applications. Springer Science & Business Media, 2012. P. 1–35.
15. *Jisa D., Thomas C.* Efficient DDoS flood attack detection using dynamic thresholding on flow-based network traffic // Computers & Security. 2019. V. 82. P. 284–295.
16. *Song Y. et al.* System level user behavior biometrics using Fisher features and Gaussian mixture models // IEEE Security and Privacy Workshops. 2013. P. 52–59.

17. *Harilal A. et al.* The Wolf Of SUTD (TWOS): A Dataset of Malicious Insider Threat Behavior Based on a Gamified Competition // J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl. 2018. V. 9, № 1. P. 54–85.
18. *Lindauer B.*Insider Threat Test Dataset. Carnegie Mellon University. Dataset. [Электронный ресурс]. URL: https://doi.org/10.1184/R1/12841247.v1 (дата обращения: 23.10.2022).
19. *Glasser J., Lindauer B.* Bridging the gap: A pragmatic approach to generating insider threat data // IEEE Security and Privacy Workshops. 2013. P. 98–104.
20. *Al-Shehari T., Alsowail R. A.* An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques // Entropy. 2021. V. 23, № 10. P. 1258.
21. *Jiang W. et al.* An insider threat detection method based on user behavior analysis //International Conference on Intelligent Information Processing. Springer, Cham, 2018. P. 421–429.
22. *Bartoszewski F. W. et al.* Anomaly Detection for Insider Threats: An Objective Comparison of Machine Learning Models and Ensembles // IFIP International Conference on ICT Systems Security and Privacy Protection. Springer, Cham, 2021. P. 367–381.
23. *Aldairi M., Karimi L., Joshi J.* A trust aware unsupervised learning approach for insider threat detection // IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). 2019. P. 89–98.
24. *Dosh M.* Detecting insider threat within institutions using CERT dataset and different ML techniques // Periodicals of Engineering and Natural Sciences. 2021. V. 9, № 2. P. 873–884.
25. *Zou S. et al.* Ensemble strategy for insider threat detection from user activity logs // Computers, Materials and Continua. 2020.
26. *Le D. C., Zincir-Heywood N., Heywood M. I.* Analyzing data granularity levels for insider threat detection using machine learning // IEEE Transactions on Network and Service Management. 2020. V. 17, № 1. P. 30–44.
27. *Ferreira P., Le D. C., Zincir-Heywood N.* Exploring feature normalization and temporal information for machine learning based insider threat detection // Proc. 15th International Conference on Network and Service Management (CNSM), 2019. P. 1–7.
28. *Мещеряков Р. В., Исхаков А. Ю., Евсютин О. О.* Современные методы обеспечения целостности данных в протоколах управления киберфизических систем // Тр. СПИИРАН. 2020. № 19 (5). С. 1089–1122.

Статья поступила в редакцию 05.11.2022;
переработанный вариант – 25.11.2022.

Исхаков Андрей Юнусович
к.т.н., старший научный сотрудник ИПУ РАН (117997, Москва, ул. Профсоюзная, д. 65), тел. (495) 198-17-20, доб. 1625, e-mail: iskhakovandrey@gmail.com.

Гайдук Кирилл
студент МИФИ (115409, Москва, Каширское шоссе, д. 31).

Using machine learning techniques for insider threat detection

Andrey Y. Iskhakov
PhD, senior researcher, ICS RAS (65 Profsoyuznaya street, Moscow 117997, Russia), e-mail: iskhakovandrey@gmail.com.

Kirill A. Gaiduk
Student, National Research Nuclear University MEPHI. Moscow Engineering Physics Institute (31 KashirskoeShosse, Moscow, 115409)

This paper presents an analysis of algorithms and approaches used to solve the problem of identifying insider threats using machine learning techniques. Internal threat detection in the context of this research is reduced to the task of detecting anomalies in the audit logs of access subjects' actions. The paper formalizes the main directions of insider threats detection and presents popular machine learning algorithms. The paper raises the problem of objective evaluation of research and development in the subject area. Based on the analysis recommendations for the implementation of internal threat detection systems using machine learning algorithms are developed.

Keywords: internal threats to information security, machine learning, anomaly hunting, authentication, isolation forest, ensemble methods.

References

1. Kim A., Oh J., Ryu J., Lee K. A Review of Insider Threat Detection Approaches with IoT Perspective. *IEEE Acces*. 2020. v. 8. pp. 78847-78867.
2. Kim J., Park M., Kim H., Cho S., Kang P. Insider Threat Detection Based on user Behavior Modeling and Anomaly Detection Algorithms. *Appl. Sci*. 2019. v. 9, 4018.
3. Alpaydin E. *Introduction to Machine Learning*. MIT Press: Cambridge, MA, 2014
4. Al-Mhiqani M.N. et al. A review of insider threat detection: Classification, machine learning techniques, datasets, open challenges, and recommendations. *Applied Sciences*. 2020. v. 10, no. 15. p. 5208.
5. Al-Mhiqani M.N. et al. A new intelligent multilayer framework for insider threat detection. *Computers & Electrical Engineering*. 2022. v. 97. p. 107597.
6. Rajaguru H., Chakravarthy S.R.S. Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. *Asian Pacific journal of cancer prevention: APJCP*. 2019. v. 20, no. 12. p. 3777.
7. Sarma M.S. et al. Insider threat detection with face recognition and KNN user classification. *2017 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*. IEEE, 2017. pp. 39-44.
8. Chauhan V.K., Dahiya K., Sharma A. Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review*. 2019. v. 52, no. 2. pp. 803-855.
9. Khan S.S., Madden M.G. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*. 2014. v. 29, no. 3. pp. 345-374.
10. Buczak A.L., Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*. 2015. v. 18, no. 2. pp. 1153-1176.
11. Le D.C., Zincir-Heywood N. Anomaly detection for insider threats using unsupervised ensembles. *IEEE Transactions on Network and Service Management*. 2021. v. 18, no. 2. pp. 1152-1164.
12. Sadaf K., Sultana J. Intrusion detection based on autoencoder and isolation forest in fog computing. *IEEE Access*. 2020. v. 8. pp. 167059-167068.
13. Hariri S., Kind M.C., Brunner R.J. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*. 2019. v. 33, no. 4. pp. 1479-1489.
14. Zhang C., Ma Y. (ed.). Ensemble machine learning: methods and applications. *Springer Science & Business Media*, 2012. pp. 1-35.
15. Jisa D., Thomas C. Efficient DDoS flood attack detection using dynamic thresholding on flow-based network traffic. *Computers & Security*. v. 82. 2019. pp. 284-295.
16. Song Y. et al. System level user behavior biometrics using Fisher features and Gaussian mixture models. *2013 IEEE Security and Privacy Workshops*. IEEE, 2013. pp. 52-59.
17. Harilal A. et al. The Wolf Of SUTD (TWOS): A Dataset of Malicious Insider Threat Behavior Based on a Gamified Competition. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl*. 2018. v. 9, no. 1. pp. 54-85.
18. Lindauer B. *Insider Threat Test Dataset*. Carnegie Mellon University. Dataset, available at: `https://doi.org/10.1184/R1/12841247.v1`(accessed 23.10.2022).
19. Glasser J., Lindauer B. Bridging the gap: A pragmatic approach to generating insider threat data. *2013 IEEE Security and Privacy Workshops*. IEEE, 2013. pp. 98-104.
20. Al-Shehari T., Alsowail R.A. An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques. *Entropy*. 2021. v. 23, no. 10. p. 1258
21. Jiang W. et al. An insider threat detection method based on user behavior analysis. *International Conference on Intelligent Information Processing*. Springer, Cham, 2018. pp. 421-429.

22. Bartoszewski F.W. et al. Anomaly Detection for Insider Threats: An Objective Comparison of Machine Learning Models and Ensembles. *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, Cham, 2021. pp. 367-381.
23. Aldairi M., Karimi L., Joshi J. A trust aware unsupervised learning approach for insider threat detection. *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2019. pp. 89-98.
24. Dosh M. Detecting insider threat within institutions using CERT dataset and different ML techniques. *Periodicals of Engineering and Natural Sciences*. 2021. v. 9, no. 2. pp. 873-884.
25. Zou S. et al. Ensemble strategy for insider threat detection from user activity logs. *Computers, Materials and Continua*. 2020.
26. Le D.C., Zincir-Heywood N., Heywood M.I. Analyzing data granularity levels for insider threat detection using machine learning. *IEEE Transactions on Network and Service Management*. 2020. v. 17, no. 1. pp. 30-44.
27. Ferreira P., Le D.C., Zincir-Heywood N. Exploring feature normalization and temporal information for machine learning based insider threat detection. *2019 15th International Conference on Network and Service Management (CNSM)*. IEEE, 2019. pp. 1-7.
28. Meshcheryakov R.V., Iskhakov A.Y., Evsyutin O.O. Sovremennye metody obespecheniya tselostnosti dannykh v protokolakh upravleniya kiberfizicheskikh system [Modern methods for ensuring data integrity in control protocols of cyber-physical systems]. *SPIIRAN proceedings*.2020. 19(5). p. 1089-1122.