Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления» (ИУ)

КАФЕДРА «Информационная безопасность» (ИУ8)

ОТЧЁТ ПО ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ

Тип практики: Преддипломная практика

Название предприятия: НУК ИУ МГТУ им. Н.Э.Баумана

Студент: группа ИУ8-124-2025 л.д. 19У113
Александров Алексей Николаевич _____ (подпись, дата)

Руководитель от предприятия:
доцент кафедры ИУ8 Коннова Наталья Сергеевна _____ (подпись, дата)

Руководитель ВКР:
доцент кафедры ИУ8 Коннова Наталья Сергеевна _____ (подпись, дата)

Руководитель от кафедры:
доцент кафедры ИУ8 Зайцева Анастасия Владленовна _____ (подпись, дата)

Оценка: _____

Москва, 2025

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления» (ИУ)

КАФЕДРА «Информационная безопасность» (ИУ8)

ИНДИВИДУАЛЬНОЕ ЗАДАНИЕ НА ПРАКТИКУ

Название предприятия: НУК ИУ МГТУ им. Н.Э.Баумана

Сроки практики: с 07.02.2025 по 07.03.2025

Специальность / направление: 10.05.01 Компьютерная безопасность

Специализация / профиль: 10.05.01_01 Математические методы защиты информации

За время прохождения практики студенту надлежит согласно программе практики:

• Сформулировать математическую постановку задачи обнаружения аномалий в сетевом трафике на основе машинного обучения.

• Предложить пути решения поставленной задачи.

Руководитель от кафедры:
доцент кафедры ИУ8 Зайцева Анастасия Владленовна _____ (подпись, дата)

Руководитель ВКР:
доцент кафедры ИУ8 Коннова Наталья Сергеевна _____ (подпись, дата)

Руководитель от предприятия:
доцент кафедры ИУ8 Коннова Наталья Сергеевна _____ (подпись, дата)

Студент: группа ИУ8-124-2025 л.д. 19У113
Александров Алексей Николаевич _____ (подпись, дата)

Является обязательным листом отчёта по практике. Документ не должен содержать информацию, отнесённую в установленном порядке к государственной тайне РФ.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современном информационном обществе огромное количество данных поступает из различных источников, таких как сети передачи данных, датчики, мобильные устройства и облачные платформы. Обработка и анализ этих данных стали неотъемлемой частью многих сфер, однако одним из серьезных вызовов является обнаружение аномалий в данных, которые могут свидетельствовать о наличии угроз безопасности или нештатных ситуациях. В этом контексте, проблема обнаружения аномалий в данных выходит на первый план. Аномалии могут служить индикаторами возможных атак, ошибок в данных или других проблем, которые могут существенно повлиять на надежность и безопасность информационных систем.

Под аномалией или выбросом (от англ. outlier) [1] понимают любой несогласованный или избыточный экземпляр выборки, отличающийся от базовой модели. Обычно это происходит, когда данные отклоняются от установленного набора данных по различным причинам, например, при неполной загрузке или неожиданном удалении информации в базе данных. Обнаружение аномалий помогает отладить процесс выявления выбросов и их устранения во избежание дефектов в наборе данных. В условиях постоянного роста объема данных, методы машинного обучения [2] обеспечивают эффективную обработку больших массивов информации. Также важным преимуществом является способность выявлять сложные и неочевидные закономерности в данных, что делает их более эффективными в обнаружении аномалий.

С появлением больших данных и широкого использования информационных технологий возросла необходимость в создании эффективных [3] средств обнаружения аномалий в данных. В зависимости от предметной области и конкретной информационной системы, аномалии могут быть признаком несанкционированного доступа, вредоносной активности или технических сбоев.

В целом, некорректные данные могут привести к неправильным аналитическим выводам, неправильным рекомендациям и даже серьезным финансовым и организационным потерям. Так недавнее исследование SAS (американская компания-разработчик систем класса Busyness Intelligence и ПО для статистического анализа) [4] в Европе показало растущую важность качества данных в финансовой сфере. 66% европейских компаний подтвердили, что ошибки в

данных негативно влияют на прибыль, 74% процента из них приняли меры по решению проблем качества данных. В связи с этим возникает потребность в совершенствовании методов анализа данных, способных оперативно выявлять нештатные ситуации.

В частности, обнаружение аномалий методами машинного обучения имеет широкий спектр применения и приложений: обнаружение инцидентов информационной безопасности по аномалиям [5], выявление неисправностей в технических системах и IoT [6], выявление мошеннической деятельности и фальсификаций [7]. Выбор методов на основе ML также связан с способностью обнаружения неявных и сложных закономерностей в данных [8], что обеспечивает эффективное выявление аномалий, которые могли бы быть упущены традиционными методами. Возможность обучаться на основе больших объемов данных, является эффективной в обнаружении новых, ранее неизвестных типов аномалий.

Место прохождения практики – Научно-учебный комплекс «Информатика и системы управления» МГТУ им. Н.Э. Баумана. Период прохождения практики – с 7 февраля 2025 г. по 7 марта 2025 г.

Цели прохождения практики – сформулировать и предложить варианты решения математической постановки задачи обнаружения аномалий в сетевом трафике на основе машинного обучения.

Задачи прохождения практики:

- Предложить формулировку математической постановки задачи обнаружения аномалий в сетевом трафике на основе машинного обучения.
- Рассмотреть, проанализировать и предложить возможные пути решения поставленной задачи.
- Представить выбор и обоснование решения математической постановки настоящей задачи.

ОСНОВНАЯ ЧАСТЬ

В рамках прохождения преддипломной практики был подготовлен настоящий отчёт о проделанной работе. Основная часть состоит из четырёх основных разделов, затрагивающих формулировку и решение математической постановки задачи обнаружения аномалий в сетевом трафике на основе машинного обучения.

Основными объектами изучения стали актуальные подходы к детектированию аномалий сетевых пакетов, выбор оптимального алгоритма решения данной проблемы, а также математический аппарат, используемый применительно к формализации и описанию рассматриваемой задачи.

В первом разделе представлена характеристика организации Научно-учебного комплекса «Информатика и системы управления» (НУК ИУ) МГТУ им. Н.Э. Баумана, где описаны цели и задачи данного подразделения, а также его роль в образовательном и научном процессах.

Во втором разделе формулируется математическая постановка задачи, где вводится множество экземпляров сетевой активности и их признаковое описание. Определяются метки классов для аномальных и нормальных экземпляров, а также формулируется функция принятия решений, позволяющая классифицировать данные. Также рассматриваются ошибки I-го и II-го рода, их взаимосвязь, позволяющую выйти на окончательную формулировку задачи.

Следующий, третий раздел посвящён обзору и описанию различных методов и подходов к решению сформулированной задачи. В этом разделе производится анализ существующих методов, таких как критерий «трёх сигм», одноклассовый метод опорных векторов, изолирующий лес и автокодировщики. В заключение, сделан выбор в пользу одного из методов, обоснованный его эффективностью и соответствием требованиям задачи. Это позволит не только продемонстрировать практическую значимость проведённой работы, но и предложить рекомендации по дальнейшему развитию и улучшению систем обнаружения аномалий в сетевом трафике.

1 Характеристика организации

Научно-учебный комплекс «Информатика и системы управления» [9] (далее — НУК ИУ) является самостоятельным структурным подразделением МГТУ им. Н.Э. Баумана (далее — Университет). Полное наименование НУК ИУ на английском языке Scientific Educational Complex «Informatics and Control Systems» Bauman Moscow State Technical University. Сокращённое — SEC ICS BMSTU.

НУК ИУ возглавляет руководитель НУК ИУ д.т.н., профессор и заведующий кафедрой ИУ6 («Компьютерные системы и сети») Пролетарский А.В., который непосредственно подчиняется действующему ректору Университета к.т.н. Гордину М.В..

Цель деятельности НУК ИУ: обеспечение образовательного и научного процессов в соответствии с Уставом Университета. К основным задачам подразделения относят:

– Подготовку бакалавров, специалистов, магистров, научных и научно-педагогических кадров высшей квалификации в соответствии с государственными лицензиями, выданными Университету.

– Выполнение фундаментальных, поисковых и прикладных научных исследований, проведение опытно-конструкторских работ и производство опытных образцов перспективной техники и технологий.

– Написание и издание учебников, учебных пособий и монографий.

– Развитие научных, педагогических и инженерных школ НУК ИУ.

– Разработку и внедрение прогрессивных форм, методов и средств подготовки специалистов.

– Повышение квалификации и переподготовка научно-педагогических кадров технических и экономических учебных заведений, а также специалистов предприятий с высшим образованием.

– Развитие материально-технической базы НУК ИУ.

– Организацию научной работы студентов.

– Развитие различных форм сотрудничества с высшими учебными заведениями России и зарубежных стран.

– Социальную поддержку и защиту преподавателей, научного и отвечающих требованиям подготовки вспомогательного персонала.

2 Математическая постановка задачи

Пусть $T = \{t_1, t_2, \ldots, t_k\}$ — множество экземпляров сетевой активности. Каждому из объектов данного множества по формуле (1) поставим в соответствие *признаковое описание*, полученное биекцией $\xi : T \to X,\ |T| = |X| = k$.

$$X = \{\overline{x_i} = \xi(t_i) \mid \forall t_i \in T\} = \{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_k}\}, \tag{1}$$

где каждый вектор $\overline{x_i},\ i = \overline{1,k}$ представляет собой набор характеристик, которые достаточно полно описывают соответствующий экземпляр t_i в рамках данной задачи.

Также соотношением (2) определим множество меток классов принадлежности экземпляров из T к аномальному трафику:

$$Y = \{y_1,\ y_2\}, \tag{2}$$

где y_1 — метка класса аномальных экземпляров (или «аномальные»);
y_2 — метка класса экземпляров без аномалий (или «нормальные»).

По формуле (3) кусочно зададим функцию принятия решений f, которая для каждого экземпляра $t_i \in T$ на основе соответствующего вектора признаков $\overline{x_i} \in X$, однозначно позволяет определить метку класса из Y:

$$f(\overline{x_i}) = \begin{cases} 1, & \text{если } \overline{x_i} \text{ соответствует классу } y_1, \\ -1, & \text{если } \overline{x_i} \text{ соответствует классу } y_2, \end{cases} \tag{3}$$

где оценку корректности функции f можно получить с помощью ошибок первого и второго рода.

Ошибкой I-го рода (α-ошибкой, ложноположительным заключением, англ. false positive) будем называть ситуацию, при которой функция f отнесёт экземпляр $t_i \in T$ к классу y_1 («аномальных»), хотя фактически он является «нормальным», то есть относится к классу y_2. Вероятность возникновения данного типа ошибки (также называемая *уровнем значимости*) определяется соотношением (4).

$$\alpha = P\Big\{ f(\overline{x_i}) = 1 \;\Big|\; t_i \ - \ \text{«нормальный» экземпляр} \Big\} \tag{4}$$

Ошибкой II-го рода (β-ошибкой, ложноотрицательное заключением, англ. false negative) будем называть ситуацию, при которой функция f отнесёт экземпляр $t_i \in T$ к классу y_2 («нормальных»), хотя фактически он является «аномальным», то есть относится к классу y_1. Вероятность возникновения данного типа ошибки по аналогии определяется соотношением (5).

$$\beta = P\Big\{ f(\overline{x_i}) = -1 \;\Big|\; t_i \;-\; \text{«аномальный» экземпляр} \Big\} \tag{5}$$

Также с этой величиной тесно связана другая, имеющая большое статистическое значение — *мощность критерия* $(1 - \beta)$. Таким образом, чем выше мощность критерия, тем меньше вероятность совершить ошибку второго рода.

Ошибки первого и второго рода взаимосвязаны между собой: при снижении уровня значимости увеличивается вероятность пропуска аномального трафика (уменьшается мощность критерия). В контексте текущей задачи, ошибка II-го рода может иметь более серьезные последствия, так как пропуск аномального трафика может привести к утечке данных или другим угрозам безопасности. В то же время, распространение ошибки I-го рода может негативно сказаться на пользовательском опыте администраторов и операторов разрабатываемой системы, поскольку это приводит к избыточной обработке и разбору ложноположительных инцидентов.

В результате задача, представленная в (6), заключается в минимизации вероятности ошибки второго рода при классификации, при этом на уровень значимости α накладывается ограничение значением порога p, которое определяет допустимый уровень ошибки I-го рода:

$$\begin{cases} \beta \to \min\limits_{\substack{\overline{x_i} \;:=\; \xi(t_i) \\ f(\overline{x_i})}}, \\ \alpha \leq p. \end{cases} \tag{6}$$

На основании исследований, представленных в работах [10], [11], [12] принято решение установить пороговое значение ошибки первого рода на уровне $p = 0,05$, что позволит обеспечить баланс между безопасностью и удобством использования системы. Это значение будет использоваться для настройки алгоритмов классификации и оптимизации их работы. Также оно будет проверено результатами проводимых численных экспериментов.

3 Пути решения поставленной задачи

В данном разделе представлены некоторые методы и подходы к решению сформулированной выше задачи.

Первый подход, который будет рассмотрен, это классический статистический метод – критерий «трёх сигм», основанный на предположении о нормальном распределении данных и позволяет выявлять аномалии, находящиеся за пределами трех стандартных отклонений от среднего значения. Несмотря на свою простоту и легкость в реализации, данный метод имеет ограничения, особенно в случаях, когда данные не подчиняются нормальному распределению.

Следующим методом является одноклассовый метод опорных векторов (One-Class SVM). Этот подход позволяет выделять области с высокой плотностью данных и эффективно обнаруживать аномалии, используя концепцию гиперплоскостей в пространстве признаков. One-Class SVM демонстрирует высокую эффективность в задачах, где аномальные данные редки и сложно выделяются.

Третий метод, который будет обсужден, это изолирующий лес (Isolation Forest). Этот алгоритм основан на принципе, что аномалии легче изолировать, так как они редки. Изолирующий лес использует случайные деревья для разделения данных и позволяет быстро и эффективно обнаруживать аномалии в больших объемах данных. В дополнение к этому, будет рассмотрено расширение данного метода — Deep Isolation Forest, которое улучшает результаты за счет использования нейронных сетей.

Также будет представлен алгоритм ECOD (Empirical-Cumulative-distribution-based Outlier Detection), который использует эмпирическую кумулятивную функцию распределения для оценки вероятности появления аномалий. Этот метод отличается простотой и отсутствием гиперпараметров, что делает его удобным для практического применения.

Наконец, в разделе будет рассмотрено применение автокодировщиков для обнаружения аномалий. Автокодировщики представляют собой нейронные сети, которые обучаются восстанавливать нормальные данные и могут быть использованы для выявления аномалий на основе ошибки восстановления. Этот подход позволяет эффективно обрабатывать сложные данные и выявлять аномалии, которые не были представлены в обучающем наборе.

3.1 Критерий «трёх сигм». Расстояние Махаланобиса

Для начала рассмотрим классический и основополагающий метод обнаружения аномалий с использованием критерия «трех сигм». Этот метод является статистическим и основывается на предположении, что данные распределены нормально, и аномалии могут быть выявлены как те, которые выходят за пределы трех среднеквадратичных отклонений от среднего значения. Из простоты метода следуют очевидные ограничения на область его применения: при значительном отклонении данных от распределения Гаусса эффективность и точность метода может значительно снизиться. Для использования критерия необходимо рассчитать среднее значение и стандартное отклонение по формулам (7) и (8) соответственно.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{7}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}, \tag{8}$$

где n – количество элементов выборки,
x_i — i-ый элемент выборки.

Затем аномалии определяются как те значения, которые находятся за пределами интервала, определенного трех среднеквадратичных отклонений от среднего (см. рисунок 1).

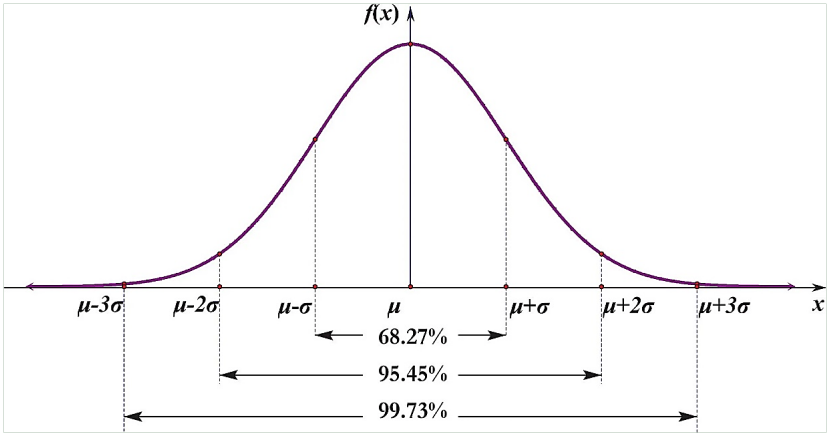

Рисунок 1 – График нормального распределения и процент попадания в отрезки, равные среднеквадратичному отклонению [13]

Помимо своей простоты, данный метод имеет ограничения в эффективности при обнаружении сложных нелинейных закономерностей в данных. Однако в многомерном случае признаки могут коррелировать между собой, что исключает прямое применение данного подхода. Для учета ковариационной структуры признаков вводится квадрат расстояния Махаланобиса, заданный соотношением (9).

$$D_M^2(\overline{x}_i) = (\overline{x}_i - \overline{\mu})^T S^{-1} (\overline{x}_i - \overline{\mu}), \tag{9}$$

где $\overline{\mu}$ — *вектор* средних значений соответствующих признаков,
S — ковариационная матрица.

Функция принятия решения в многомерном случае принимает вид (10):

$$f(\overline{x}_i) = \begin{cases} 1, & \text{если } D_M(\overline{x}_i) > r, \\ -1, & \text{иначе,} \end{cases} \tag{10}$$

где порог r подбирается так, чтобы вероятность ложноположительного срабатывания (ошибка I рода) не превышала заданное значение.

При условии, что вектор признаков $\overline{x}_i$ распределён нормально, величина $D_M^2(\overline{x}_i)$ следует распределению χ_d^2 с d степенями свободы (где d — размерность вектора признаков), порог r можно определить из равенства (11).

$$P\Big\{\chi_d^2 > r^2\Big\} = p \tag{11}$$

Тогда вероятности ошибок I-го и II-го рода определяются соотношениями (12) и (13) соответственно.

$$\alpha = P\Big\{D_M(\overline{x}_i) > r \;\Big|\; t_i \;-\; \text{«нормальный» экземпляр}\Big\}. \tag{12}$$

$$\beta = P\Big\{D_M(\overline{x}_i) \leq r \;\Big|\; t_i \;-\; \text{«аномальный» экземпляр}\Big\}. \tag{13}$$

Таким образом, метод позволяет установить строгие границы для функции принятия решения f. При условии нормальности распределения данных, критерий обеспечивает очень малую вероятность ошибки I-го рода, что удовлетворяет ограничению, заданному в постановке задачи, и позволяет эффективно минимизировать β, что соответствует цели оптимизации, сформулированной в

(6). Однако сетевой трафик редко может быть описан распределением Гаусса, ввиду нерегулярности событий в вычислительных сетях, что даёт повод для рассмотрения других путей решения задачи. Кроме того, такой подход не учитывает структуру данных и контекст анализа. Однако он включён рассмотрение ввиду его первоначальной каноничности для решения поставленной задачи. Многие статистические модели используют, развивают и усовершенствуют данный подход.

3.2 Одноклассовый метод опорных векторов, One-Class SVM

One-Class Support Vector Machine (One-Class SVM или просто One-SVM) представляет собой одноклассовый неконтролируемый метод обнаружения аномалий [14]. Основной идеей One-SVM является установление границы, охватывающей области с высокой плотностью данных, при этом исключая заданную долю данных $\nu = (0;\ 1]$ (гиперпараметр модели), которая (по мнению пользователя модели) представляет собой верхнюю границу числа аномалий в данных (см. рисунок 2).

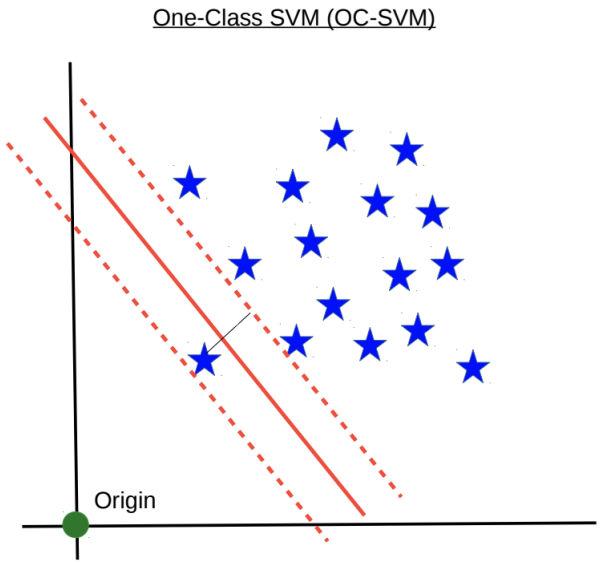

Рисунок 2 – Специализация метода опорных векторов для поиска выбросов [15]

Главной целью One-SVM является поиск гиперплоскости, разделяющей данные от источника (начала координат). Для достижения этой цели метод использует линейный классификатор, но с использованием ядра — функции, которая отображает пространство признаков в пространство высшей размерности. Канонично данный kernel trick [16] позволяет разделять классы, линейно неразделимые в текущем признаковом пространстве (см. рисунок 3). Таким образом на выходе получается модель с нелинейными границами в исходном пространстве переменных задачи.

RBF (Radial Basis Function) [17] часто является наиболее широко применяемым ядром для One-SVM ввиду нелинейности и гибкой адаптации к различным распределениям данных. Данное ядро определяется функцией (14):

$$K_{RBF}(\overline{x},\ \overline{x}') = e^{-\gamma\|\overline{x}-\overline{x}'\|^2}, \tag{14}$$

где $\overline{x},\ \overline{x}' \in \mathbb{R}^k$ — выборки, векторы в входном пространстве размерности k, γ — свободный параметр, который позволяет настраивать влияние соседних точек на границу принятия решения.

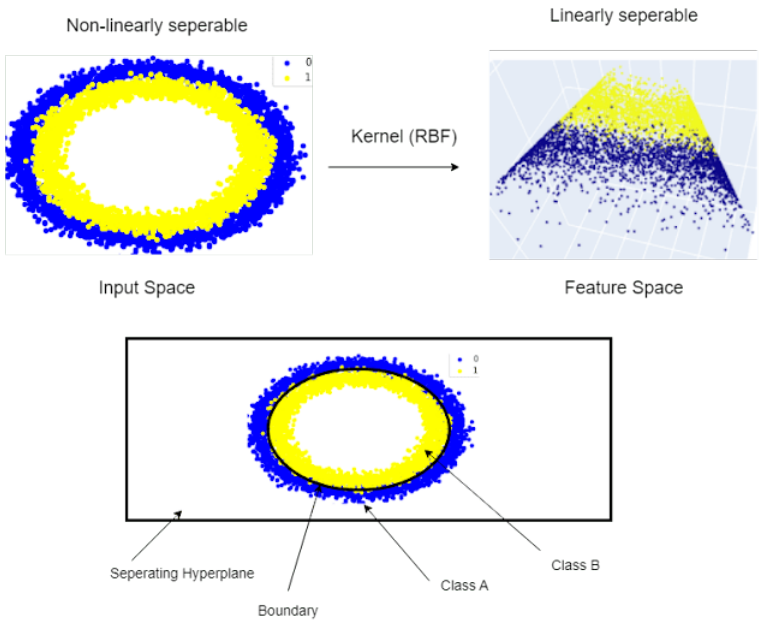

Рисунок 3 – Нелинейное разделение данных на основе пороговой гиперплоскости в пространстве признаков [17]

Таким образом, в результате обучения One-Class SVM стремится создать гиперплоскость, которая наилучшим образом ограничит область с высокой плотностью данных (см. рисунок 4), считая все остальные области аномальными. Этот метод демонстрирует свою эффективность в выявлении аномалий и обеспечивает контроль над тем, какая доля данных считается аномальной, благодаря гиперпараметру ν.

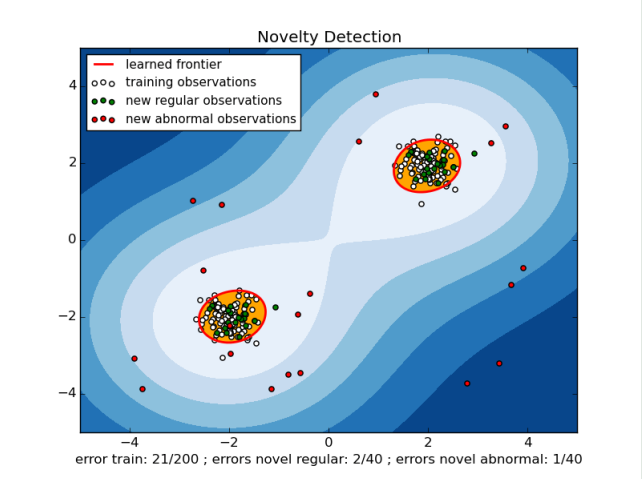

Рисунок 4 – Пример обнаружения аномалий на основе One-SVM с нелинейным ядром RBF [18]

Оптимизация модели One-Class SVM направлена на минимизацию β при условии, что вероятность ложноположительного срабатывания α не превышает установленный порог p. Применение нелинейного ядра (например, RBF), позволяет отобразить исходное пространство признаков в пространство более высокой размерности, что способствует построению адаптивной и нелинейной границы принятия решения, способной учитывать сложную структуру сетевого трафика. Это непосредственно соответствует цели оптимизации, сформулированной в (6).

3.3 Изолирующий лес. Расширение Deep IForest

Изолирующий лес (Isolation Forest или IForest) — это метод машинного обучения без учителя, основанный на принципе, что аномалии обычно редки и хорошо выделяются, что упрощает их идентификацию. IForest основан на идее разделения пространства признаков для обнаружения аномалий. Однако, в отличие от деревьев решений, где разделение основано на информации, в IForest процесс разделения является рандомизированным.

Метод IForest строит деревья [19], выбирая случайным образом значения для разделения между минимальным и максимальным значениями функции, которая также выбирается случайным образом из предварительно заданного набора функций. Эти случайные разделения создают более короткие пути в деревьях для аномальных наблюдений, отделяя их от общего множества данных (см. рисунок 5). Аномалии требуют меньшего разделения, потому что плотность вокруг них низкая.

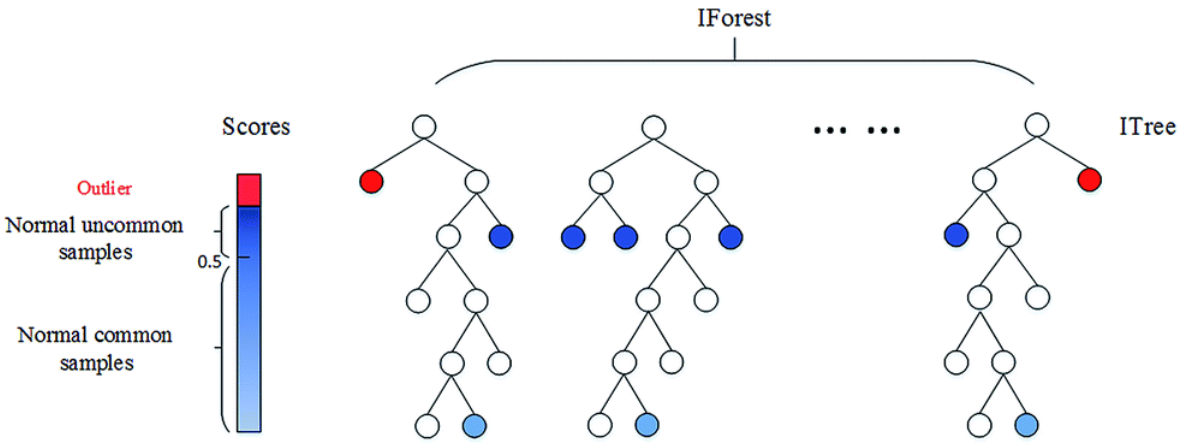

Рисунок 5 – Схема выявления аномалий на основе метода изолирующего леса [19]

Основная идея IForest заключается в том, чтобы вычислить оценку аномалии для каждого наблюдения в наборе данных. Для этого строится лес случайных и независимых бинарных изолирующих деревьев (ITree). Также модель использует два входных параметра: размер выборки ψ, случайно выбранный из всего набора данных, и t — количество деревьев в лесу.

Построение очередного ITree происходит следующим образом:

– инвариант – корневой узел содержит все наблюдения выборки;
– IForest случайным образом выбирает подмножество наблюдений ψ;
– разделение узла i:
  - IForest случайным образом выбирает функцию f_i из заданного набора функций;
  - значение разделения v_i также случайным образом выбирается между $\min(f_i)$ и $\max(f_i)$;
  - элементы узла i разбиваются на левую и правую группы (поддеревья) путем сравнения их значений с v_i;
– процедура разделения рекурсивно продолжается до тех пор, пока все выборочные данные не будут изолированы или дерево не достигнет максимальной глубины, равной $\log_2 \psi$.

Так, чтобы построить t изолирующих деревьев леса, необходимо повторить вышеописанные шаги t раз. Построение IForest представляет собой этап обучения в данном методе.

Для каждого наблюдения можно рассчитать показатель выбросов [20] по формуле (15):

$$s(\overline{x},\ n) = 2^{-\frac{\mu(h(x))}{c(n)}}, \tag{15}$$

где $h(\overline{x})$ — число шагов до полной изоляции экземпляра $\overline{x}$,
$\mu(h(\overline{x}))$ — среднее значение $h(\overline{x})$, взятое по ансамблю деревьев,
$c(n)$ — значение нормализации, заданное соотношением (16).

$$c(n) = \begin{cases} 2H(m-1) - \dfrac{2(m-1)}{n}, & \text{если } m > 2, \\ 1, & \text{если } m = 2, \\ 0, & \text{иначе,} \end{cases} \tag{16}$$

где n — объем данных тестирования,
m — размер набора выборок,
$H(i)$ — число гармоник, которое можно оценить $H(i) = \ln i + \gamma$, где $\gamma \approx 0,577$ — постоянная Эйлера-Макерони [21].

Таким образом, если значение $s(\overline{x}, n)$ близко к 1, то экземпляр $\overline{x}$ с большой вероятностью является аномалией. Если $s(\overline{x}, n) < \frac{1}{2}$, то $\overline{x}$, скорее всего, относится к нормальной точке выборки.

Данный метод эффективен для обработки больших объемов данных и обладает высокой скоростью обучения, что делает его привлекательным для решения задач обнаружения аномалий в реальном времени. Однако стоит заметить, что в случае локальных выбросов IForest может справляться хуже, чем с глобальными аномалиями.

Более того, в 2023 году в работе [22] было представлено новое расширение — Deep Isolation Forest (DIF). Метод предлагает ряд существенных преимуществ по сравнению со стандартным методом IForest:

- эффективность в пространствах высокой размерности;
- избежание алгоритмического смещения (см. рисунок 6): DIF предоставляет более обширный метод изоляции, который может произвольно разделять данные в любом случайном направлении и угле на подпространствах любого размера;
- возможность обработки сложных и трудноизолируемых аномалий за счёт использования случайных нейронных сетей, не требующих оптимизации.

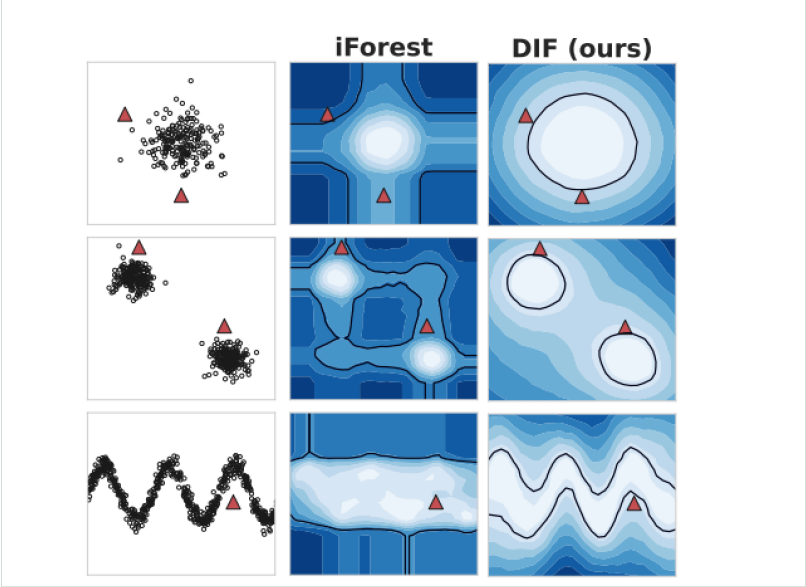

Рисунок 6 – Сравнение IForest и DIF: разрешение проблемы алгоритмического смещения при выявлении аномалий [22]

Таким образом, DIF преодолевает ограничения стандартного IForest, дополняя каноничный метод и предоставляя более гибкое и эффективное средство обнаружения аномалий.

Здесь функция принятия решений $f(\overline{x_i})$ основывается на оценке аномальности $s(\overline{x_i}, n)$, рассчитанной по формуле (15). Для определения принадлежности экземпляра $\overline{x_i}$ к классу y_1 или y_2 необходимо установить пороговое значение s^*, при превышении которого экземпляр считается аномальным. Таким образом, функция $f(\overline{x_i})$ определяется следующим образом:

$$f(\overline{x_i}) = \begin{cases} 1, & \text{если } s(\overline{x_i}, n) \geq s^*, \\ -1, & \text{иначе.} \end{cases} \tag{17}$$

Выбор порогового значения s^* напрямую влияет на α и β. Для достижения оптимального баланса между α и β необходимо производить калибровку s^* с использованием кросс-валидации.

3.4 ECOD

В марте 2022 года в работе [23] был показан простой, но эффективный алгоритм ECOD (Empirical-Cumulative-distribution-based Outlier Detection), который был вдохновлен тем фактом, что аномалии часто являются «редкими событиями», которые появляются на краях распределения. Данный подход имеет ряд важнейших особенностей, которые стоит отметить:

- простота понимания и интерпретации метода;
- отсутствие гиперпараметров, что может быть важно в контексте обнаружения аномалий: их часто может быть сложно корректно настроить, поскольку аномальные экземпляры редки и их не всегда просто получить;
- в данном методе временная сложность линейно зависит от размера входных данных и числа измерений.

ECOD использует информацию о распределении данных, чтобы определить, где данные менее вероятны и, следовательно, более вероятны выбросы. В частности, ECOD оценивает [24] эмпирическую кумулятивную функцию распределения ECDF (см. рисунок 7) для каждой переменной отдельно.

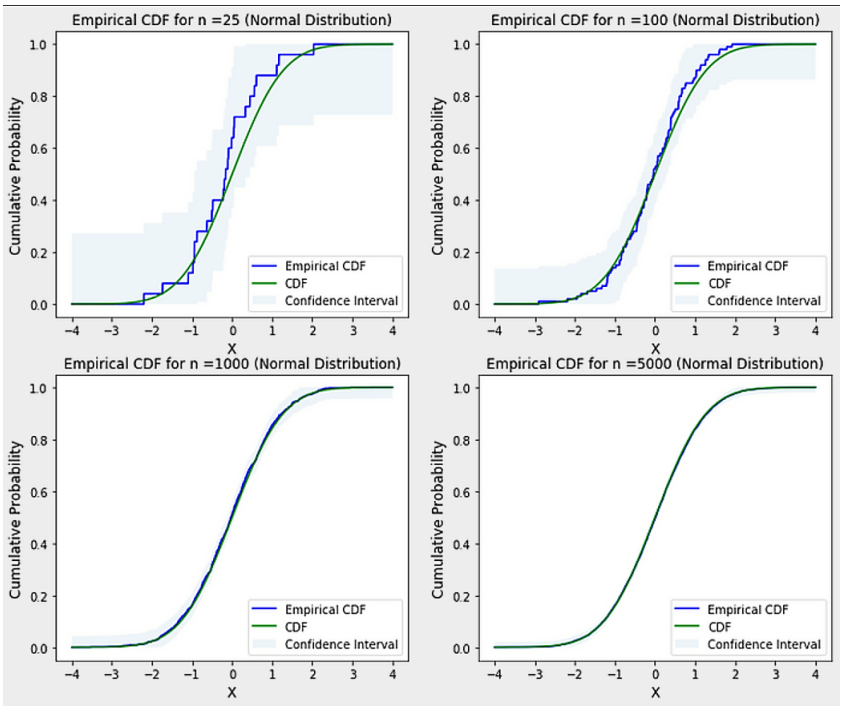

Рисунок 7 – Примеры ECDF для разного объёма выборки нормального распределения [24]

Идеальным вариантом здесь было бы применить объединенный ECDF ко всем переменным; однако это слишком дорого с точки зрения вычислений (скорость сходимости при оценке совместного CDF замедляется по мере увеличения числа переменных). Таким образом, ECOD делает упрощающее предположение о том, что переменные статистически независимы. Но даже с учетом этого предположения ECOD работает довольно хорошо.

Рассмотрим, как работает обучение в данном методе. Пусть на вход поданы данные $X = \{X_i\}_{i=1}^{n} \in \mathbb{R}^{(n \times d)}$ с n выборками и d признаками. Обозначим $X_i^{(j)}$ значение j-го признака для i-го экземпляра. Тогда для того, чтобы определить, является ли экземпляр X_i аномалией, необходимо рассчитать 3 вероятностные величины на «хвостах» (tails) распределений по формулам (18), (19) и (20).

$$O_{left}(X_i) = -\sum_{j=1}^{d} \log\left[ECDF_{left}^{(j)}(X_i^{(j)})\right], \tag{18}$$

$$O_{right}(X_i) = -\sum_{j=1}^{d} \log\left[ECDF_{right}^{(j)}(X_i^{(j)})\right], \tag{19}$$

$$O_{auto}(X_i) = \begin{cases} -\sum\limits_{j=1}^{d} \log\left[ECDF_{left}^{(j)}(X_i^{(j)})\right], & \text{если } \gamma_j < 0, \\ -\sum\limits_{j=1}^{d} \log\left[ECDF_{right}^{(j)}(X_i^{(j)})\right], & \text{иначе,} \end{cases} \tag{20}$$

где γ_j – коэффициент асимметрии (skewness) выборки для j-го распределения признака (для одной переменной нахождение в левом хвосте распределения вероятностей может быть более аномальным, чем для другой), который, в свою очередь, определяется соотношением (21).

$$\gamma_j = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i^{(j)} - \overline{X^{(j)}}\right)^3}{\left[\frac{1}{n-1}\sum\limits_{i=1}^{n}\left(X_i^{(j)} - \overline{X^{(j)}}\right)^2\right]^{3/2}}. \tag{21}$$

Одно из преимуществ использования отрицательных логарифмов вероятностей заключается в том, что более низкая вероятность преобразуется в более высокое отрицательное логарифмическое значение. Таким образом, редкие элементы с низкой вероятностью получают более высокие оценки аномалий. Тогда результирующую оценку аномальности экземпляра X_i определяется соотношением (22).

$$O(X_i) = \max\{O_{left}(X_i),\ O_{right}(X_i),\ O_{auto}(X_i)\} \tag{22}$$

На рисунке 8 показано, как комплексный подход к оценке и корректировка асимметрии на основе γ_j даёт возможность значительно улучшить результаты в выявлении аномалий в данных. Предложенный метод является эффективным, быстрым и масштабируемым алгоритмом обнаружения аномалий в данных.

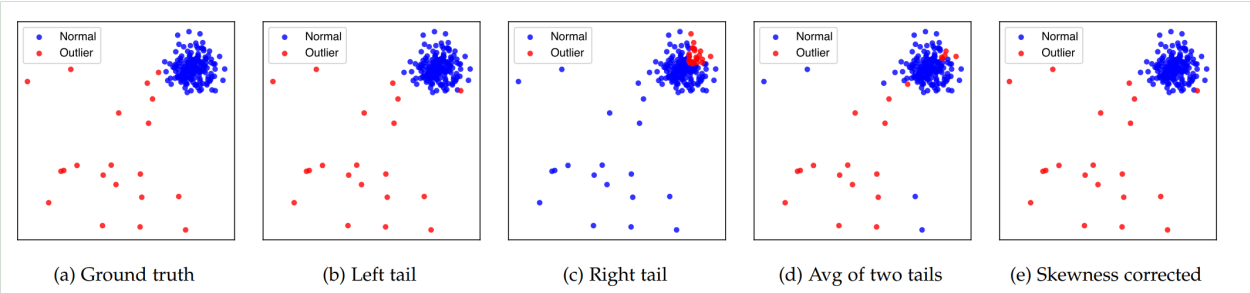

Рисунок 8 – Демонстрация влияния «хвостовых вероятностей» и корректировки асимметрии на выявление аномалий [23]

Функция принятия решений (24) формализуется через задание порогового значения $\tau \in \mathbb{R}^+$, определяемого на основе концепции уровня контаминации (англ. *contamination rate*) $\rho \in [0, 1]$, который задаёт ожидаемую долю аномалий в данных. Для этого:

1. все экземпляры ранжируются по убыванию оценок $O(X_i)$;
2. τ устанавливается в (23) как $(1-\rho)$-квантиль полученного распределения оценок:

$$\tau = Q_{1-\rho}\left(\{O(\overline{x_j}) \mid \forall \overline{x_j} \in X\}\right), \tag{23}$$

где $Q_{1-\rho}$ — оператор квантиля уровня $(1-\rho)$.

Сама функция:

$$f(\overline{x_i}) = \begin{cases} 1, & \text{если } O(\overline{x_i}) \geq \tau, \\ -1, & \text{иначе.} \end{cases} \tag{24}$$

Данный подход соответствует описанию метода ECOD в оригинальной работе [23], где:

- оценки $O(X_i)$ вычисляются независимо для каждого измерения через эмпирические хвостовые вероятности (уравнения 18–20);
- порог τ адаптируется к данным через квантильный критерий, что устраняет необходимость явных предположений о распределении нормальных экземпляров;
- уровень значимости α косвенно контролируется через параметр ρ, который может быть настроен в соответствии с требованиями задачи.

Минимизация β достигается за счёт следующих свойств ECOD:

- использование логарифмических преобразований хвостовых вероятностей усиливает контраст между типичными и редкими событиями;
- адаптивный выбор хвоста через коэффициент асимметрии γ_j (формула 21) улучшает чувствительность к мультимодальным распределениям;
- агрегация через max-оператор фокусируется на наиболее аномальных измерениях.

3.5 Автокодировщик для обнаружения аномалий

Также для решения задачи обнаружения аномалий было рассмотрено применение автокодировщиков (или автоэнкодеров, англ. autoencoder). Представляет собой архитектурный класс искусственных нейронных сетей, которая использует метод обучения без учителя с применением алгоритма обратного распространения ошибки.

В применении автокодировщиков для обнаружения аномалий используется их основное свойство — необходимость фиксировать в скрытом латентном слое, полученном с помощью кодера (англ. encoder), наиболее важной информации о входном сигнале для его последующего восстановления декодером (англ. decoder). Благодаря этому свойству автокодировщик способен восстанавливать знакомый ему сигнал, считающийся нормальным, и не может восстановить аномальный сигнал, содержащий неизвестные паттерны. Это свойство и заложено в основе применения автокодировщиков для обнаружения аномалий.

Рассмотрим нейронную сеть автокодировщика с единственным скрытым слоем. Она будет иметь кодер и декодер, заданные соотношениями (25) и (26), соответственно.

$$\overline{h} = \sigma(W_{xh}\overline{x} + \overline{b_{xh}}), \tag{25}$$

$$\overline{z} = \sigma(W_{hx}\overline{h} + \overline{b_{hx}}), \tag{26}$$

где W и $\overline{b}$ — вес и смещение (англ. bias) нейронной сети,
σ — функция нелинейного преобразования.

Кодер в уравнении (25) отображает входной вектор $\overline{x}$ в скрытое представление $\overline{h}$ с помощью нелинейного преобразования. Декодер в (26) отображает скрытое представление $\overline{h}$ обратно в $\overline{z}$ исходного пространство с помощью того же преобразования, что и кодер. Разница между исходным входным и восстановленным векторами называется ошибкой восстановления, как показано в формуле (27).

$$e = \|\overline{x} - \overline{z}\|. \tag{27}$$

Автокодировщик обучатся минимизировать e. Алгоритм обучения нативного автоэнкодера показан в алгоритме 1, где f_ϕ и g_θ - многослойные нейронные сети для автокодировщика.

Алгоритм 1 — Стандартное обучение автокодировщика

Исходные параметры: Набор данных $\overline{x_1}, ..., \overline{x_n},\ n < k$
Результат: обученные f_ϕ и g_θ

```
Инициализировать параметры ϕ и θ
до тех пор, пока не достигнута сходимость ϕ и θ выполнять
    // Вычисление суммарной ошибки восстановления.
    E = Σ_{i=1}^{n} ||x̄_i − g_θ(f_ϕ(x̄_i))||
    ϕ, θ ← Обновление параметров с помощью градиента E
      (например, стохастический градиентный спуск)
конец
возвратить f_ϕ, g_θ
```

Обнаружение аномалий на основе автокодировщика — это метод обнаружения аномалий на основе отклонений, использующий обучение с частичным привлечением учителя (англ. semi-supervised learning). Он использует ошибку восстановления в качестве оценки аномалии. Точки данных с высокой значением e считаются аномалиями.

Для обучения автоэнкодера используются только данные с нормальными экземплярами. После обучения автоэнкодер будет очень хорошо восстанавливать нормальные данные, но не сможет сделать этого с аномальными данными, с которыми автоэнкодер не сталкивался. Алгоритм 2 демонстрирует обнаружение аномалий с использованием ошибок восстановления автокодировщиков.

Алгоритм 2 — Обнаружение аномалий на основе автокодировщика

Исходные параметры:
Набор **нормальных** данных $X' \subset X,\ |X'| = n < k$,
Набор неразмеченных данных (потенциально аномальных) $\overline{x'_1}, ..., \overline{x'_l}$,
Метки классов $Y = \{y_1,\ y_2\}$, где y_1 — «аномальный»,
y_2 — «нормальный»,
Порог ошибки ϵ.
Результат: значение ошибки восстановления $e = \|\overline{x} - \overline{z}\|$

$\phi,\ \theta \leftarrow$ обучить, используя набор «нормальных» данных X'
цикл $i := \overline{1,\ l}$ **выполнять**
  $error(i) = \|\overline{x_i} - g_\theta(f_\phi(\overline{x_i}))\|$
  если $error(i) > \epsilon$ **тогда**
    $\overline{x_i}$ помечается, как принадлежащий классу y_1
  иначе
    $\overline{x_i}$ помечается, как принадлежащий классу y_2
  конец
конец

Рассмотрим архитектуру нейронной сети автокодировщика, представленной на рисунке 9, которая может быть применена для обнаружения аномалий сетевого трафика.

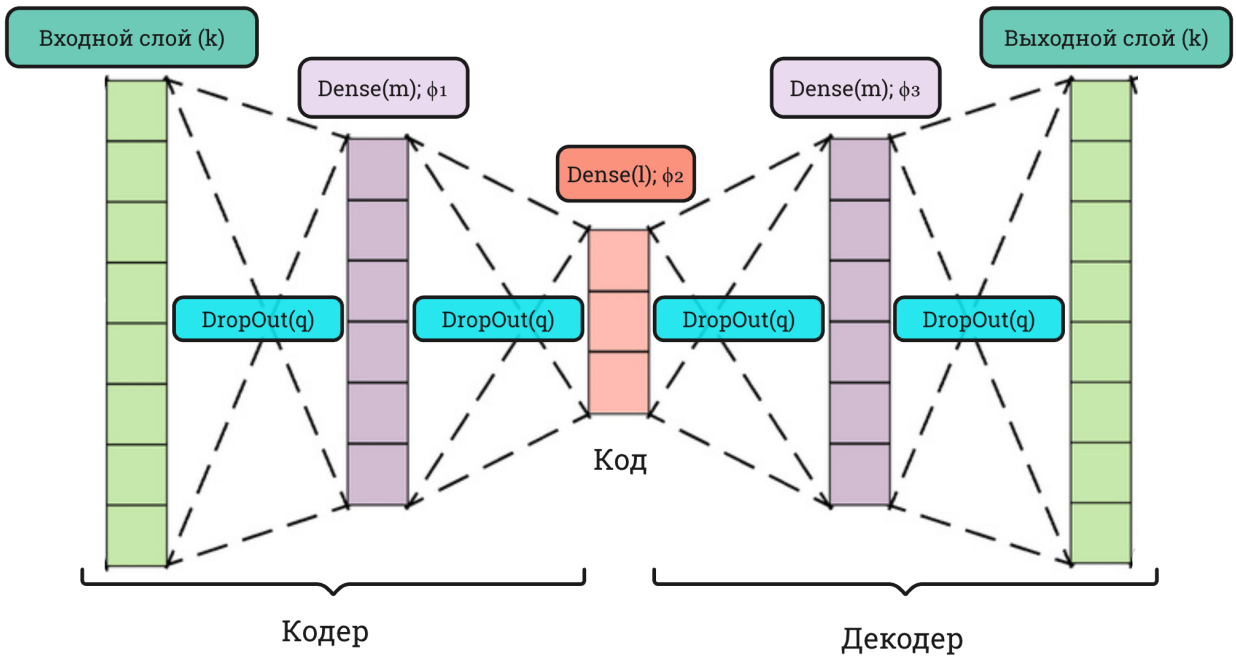

Рисунок 9 – Архитектура искусственной нейронной сети автокодировщика для обнаружения аномалий в сетевом трафике

Входной и выходной слои имеют количество нейронов по числу признаков, извлечённых в элементы X из множества T по формуле (1). Скрытые слои энкодера и декодера содержат m и l нейронов соответственно.

Каждый слой использует функции активации ϕ_1, ϕ_2 и ϕ_3 для введения нелинейности в модель. Примером хорошо зарекомендовавшей себя функции активации может послужить линейный выпрямитель (или полулинейный элемент, англ. *Rectified linear unit*, ReLU), представленный формулой (28). Она монотонна, что гарантированно делает выпуклой поверхность ошибок, ассоциированную с одноуровневой моделью. Также ReLU — гладкая, то есть имеет монотонную производную, что в некоторых случаях обеспечивает более высокую степень общности.

$$\phi(x) = \begin{cases} 0, & \text{если } x < 0, \\ x, & \text{если } x \geq 0. \end{cases} \tag{28}$$

Также в автокодировщике применяется метод регуляризации *Dropout* с параметром $q \in (0;\ 1)$, что помогает предотвратить переобучение, случайным образом отключая $q \cdot 100\%$ нейронов во время обучения.

Обучение автокодировщика может осуществляться с использованием функции потерь MSE (англ. *Mean Squared Error* — среднеквадратичная ошибка), которая измеряет разницу между входными данными и восстановленными данными. Далее результаты восстановления декодера сравниваются с входным слоем для оценки аномальности экземпляра данных. Среднеквадратичная ошибка определяется соотношением $MSE = \sigma^2$, где σ — стандартное отклонение, ранее определённое в формуле (8).

Свяжем ошибку восстановления e с функцией принятия решений $f(\overline{x_i})$ через пороговое значение ошибки восстановления $\epsilon \in \mathbb{R}^+$ по формуле (29).

$$f(\overline{x_i}) = \begin{cases} 1, & \text{если } e(\overline{x_i}) \geq \epsilon, \\ -1, & \text{иначе}, \end{cases} \tag{29}$$

где $e(\overline{x_i}) = \|\overline{x_i} - g_\theta(f_\phi(\overline{x_i}))\|$ — норма разности входного и восстановленного векторов (см. алгоритм 2).

4 Обоснование решения математической постановки задачи

В конечном итоге, обнаружение аномалий в сетевом трафике решено построить на базе автокодировщиков. Этот выбор обусловлен успешностью применения данного класса моделей для решения поставленной задачи и оптимизации параметра β при установленном значении порога α, что показано в результатах работ [25], [26], [27]. Также это подтвердили численные эксперименты, проводимые в рамках прошедших НИРС.

Автокодировщики обладают способностью эффективно обрабатывать высокоразмерные данные, что является важным аспектом при анализе сетевого трафика, где количество признаков может быть значительным. Они способны выявлять сложные паттерны в данных благодаря своей архитектуре. Это позволяет им адаптироваться к различным типам данных и выявлять аномалии, которые могут быть неочевидны для других методов.

Также данный метод позволяет осуществить тонкую настройку за счёт установки параметров скрытых полносвязных слоёв нейросети автоэнкодера, выбора оптимальных функций активации, применения метода регуляризации *Dropout* с подбором процента исключения случайных нейронов, что повышает итоговое качество предсказаний модели, а также подбор функции потерь и оптимизатора нейронной сети.

ЗАКЛЮЧЕНИЕ

В данной работе была сформулирована математическая постановка задачи обнаружения аномалий в сетевом трафике на основе машинного обучения. Были предложены пути решения поставленной задачи, выбран и описан оптимальный алгоритм для её решения. В результате решение настоящей задачи будет построено на базе нейронной сети автокодировщика, хорошо зарекомендовавшего себя по результатам опубликованных работ и численных экспериментов, проводимых ранее в рамках НИРС.

Специфика предметной области и сам тип аномалий представляет особые вызовы для существующих методов их обнаружения. На основе анализа проблемы, технических и функциональных особенностей, реализуемых в информационной системе, делается выбор в пользу тех или иных методов решения поставленной задачи. Действительно, вредоносная активность имеет тенденцию выглядеть нерегулярно по сравнению с повседневными операциями, что позволяет свести данную проблему к задаче поиска аномалий в данных сетевого трафика, поступающего как из сети Интернет, так и циркулирующего в контуре организации.

Системы, основанные на заранее настроенных сигнатурах и правилах, зачастую будут «слепы» к атакам нулевого дня, в то время как anomaly-based-решения способны выявить подозрительную активность на основе алгоритмов машинного обучения, что будет отражено в выпускной квалификационной работе специалиста.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Что такое обнаружение аномалий? Описание обнаружения аномалий с помощью машинного обучения [Электронный ресурс]. — Amazon Web Services, 2022. — URL: https://aws.amazon.com/ru/what-is/anomaly-detection/ (дата обр. 15.02.2025).

2 Обнаружение аномалий: выявление выбросов в данных [Электронный ресурс]. — wedx.ru. — URL: https://wedx.ru/obnaruzhenie-anomalij-vyyavlenie-vybrosov-v-dannyh.html (дата обр. 15.02.2025).

3 Artificial Intelligence and Machine Learning for Anomaly Detection [Электронный ресурс]. — Veritis.com, 2022. — URL: https://www.veritis.com/blog/anomaly-detection-using-machine-learning/ (дата обр. 15.02.2025).

4 SAS Data Quality and SAS Data Governance [Электронный ресурс]. — SAS: Analytics, Artificial Intelligence, Data Management, 2019. — URL: https://www.sas.com/content/dam/SAS/documents/event-collateral/2019/ru/seminar-kz-data-laboratory/seminar-kz-sas-data-quality-data-governance.pdf (дата обр. 15.02.2025).

5 *Есипов Д. А.*, *Шабала Е. Е.*, *Щетинин Б. С.* Метод обнаружения инцидентов информационной безопасности по аномалиям в биометрических поведенческих чертах пользователя. — Научно-технический вестник информационных технологий, механики и оптики, 2022. — URL: https://cyberleninka.ru/article/n/metod-obnaruzheniya-intsidentov-informatsionnoy-bezopasnosti-po-anomaliyam-v-biometricheskih-povedencheskih-chertah-polzovatelya (дата обр. 15.02.2025).

6 Методы для обнаружения и диагностика неисправностей применительно IoT [Электронный ресурс]. — Хабр, 2019. — URL: https://habr.com/ru/companies/otus/articles/567574/ (дата обр. 15.02.2025).

7 *Хрипунов П. В.*, *Минаев Е. Ю.*, *Проценко В. И.* Обнаружение аномалий и фальсификаций в данных социальных сервисов в рамках цифровой экономики. — Известия Самарского научного центра РАН, 2019. — URL: https://cyberleninka.ru/article/n/obnaruzhenie-anomaliy-i-falsifikatsiy-v-dannyh-sotsialnyh-servisov-v-ramkah-tsifrovoy-ekonomiki (дата обр. 15.02.2025).

8 Machine learning (ML) powered anomaly detection [Электронный ресурс]. — Learn Netdata, 2023. — URL: https://learn.netdata.cloud/docs/ml-and-troubleshooting/machine-learning-ml-powered-anomaly-detection (дата обр. 15.02.2025).

9 *ПОЛОЖЕНИЕ №03.02-06/2.* О структурном подразделении Научно-учебный комплекс «Информатика и системы управления». — Взамен Положения о структурном подразделении №03.02-06/1 от 12.04.2012. — МГТУ им. Н.Э. Баумана, 2022. — 8 с.

10 *Марчук В. И.*, *Румянцев К. Е. а.* Анализ методов адаптации порогового значения при обнаружении аномальных измерений. — Известия вузов России. Радиоэлектроника, 2006. — URL: https://cyberleninka.ru/article/n/analiz-metodov-adaptatsii-porogovogo-znacheniya-pri-obnaruzhenii-anomalnyh-izmereniy (дата обр. 21.02.2025).

11 *Васильев В. И.*, *Вульфин А. М.*, *Гвоздев В. Е.* Обеспечение информационной безопасности киберфизических объектов на основе прогнозирования и обнаружения аномалий их состояния // Системы управления, связи и безопасности. — 2021. — № 6. — С. 90—119. — ISSN 2410-9916. — (Дата обр. 17.02.2025).

12 *Романов Н. А.*, *Кропотов Д. А.* Детектирование аномалий во временных рядах при помощи глубоких нейронных сетей. — Московский государственный университет имени М.В. Ломоносова, 2018. — URL: http://www.machinelearning.ru/wiki/images/6/6e/2018_617_RomanovNA.pdf (дата обр. 18.02.2025).

13 *Kenton W.* Three Sigma Limits Statistical Calculation With Example. — Investopedia.com, 2024. — URL: https://www.investopedia.com/terms/t/three-sigma-limits.asp (дата обр. 20.02.2025).

14 Документация метода sklearn.svm.OneClassSVM [Электронный ресурс]. — Scikit-learn – Machine Learning in Python. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html (дата обр. 20.02.2025).

15 *Rai P.* Introduction to Machine Learning: SVM (Contd), Multiclass and One-Class SVM. — Московский государственный университет имени М.В. Ломоносова, 2018. — URL: https://www.cse.iitk.ac.in/users/piyush/courses/ml_autumn18/material/771_A18_lec11.pdf (дата обр. 20.02.2025).

16 Поиск аномалий с One-Class SVM [Электронный ресурс]. — Otus.ru – образовательная онлайн-платформа, 2019. — URL: https://otus.ru/nest/post/888/ (дата обр. 20.02.2025).

17 The RBF kernel in SVM: A Complete Guide [Электронный ресурс]. — PyCodeMates, 2022. — URL: https://www.pycodemates.com/2022/10/the-rbf-kernel-in-svm-complete-guide.html (дата обр. 20.02.2025).

18 One-class SVM with non-linear kernel (RBF) [Электронный ресурс]. — Scikit-learn – Machine Learning in Python. — URL: https://scikit-learn.sourceforge.net/dev/auto_examples/svm/plot_oneclass.html (дата обр. 20.02.2025).

19 Краткий обзор методов обнаружения выбросов [Электронный ресурс]. — Skine.ru. — URL: https://skine.ru/articles/476079/ (дата обр. 20.02.2025).

20 Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou Isolation Forest [Электронный ресурс]. — Eighth IEEE International Conference on Data Mining, 2008. — URL: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest (дата обр. 20.02.2025).

21 *Радевич В. С.* Коэффициент и постоянная Эйлера-Маскерони. — Современные инновации, 2022. — URL: https://cyberleninka.ru/article/n/koefitsient-i-postoyannaya-eylera-maskeroni (дата обр. 20.02.2025).

22 *Xu H.*, *Pang G.*, *Wang Y.* Deep Isolation Forest for Anomaly Detection. — 2023. — arXiv: 2206.06602 [cs.LG]. — URL: https://arxiv.org/pdf/2206.06602.pdf.

23 *Li Z.*, *Zhao Y.*, *Hu X.* ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. — 2022. — arXiv: 2201.00382v3 [cs.LG]. — URL: https://arxiv.org/abs/2201.00382v3.pdf.

24 Replace Outlier Detection by Simple Statistics with ECOD [Электронный ресурс]. — Medium – Where good ideas find you, 2022. — URL: https://medium.com/geekculture/replace-outlier-detection-by-simple-statistics-with-ecod-f95a7d982f79 (дата обр. 20.02.2025).

25 *Gurina A.*, *Eliseev V.* Anomaly-Based Method for Detecting Multiple Classes of Network Attacks // Inf. — 2019. — Т. 10. — С. 84. — URL: https://api.semanticscholar.org/CorpusID:86701180.

26 *Чернышов Ю. Ю.* Применение автокодировщиков для выявления аномалий в киберфизических системах. — Вестник Пермского университета, 2022. — URL: https://cyberleninka.ru/article/n/primenenie-avtokodirovschikov-dlya-vyyavleniya-anomaliy-v-kiberfizicheskih-sistemah (дата обр. 21.02.2025).

27 *Гурина А. О.*, *Гузев О. Ю.*, *Елисеев В. Л.* Обнаружение аномальных событий на хосте с использованием автокодировщика. — International Journal of Open Information Technologies, 2020. — URL: https://cyberleninka.ru/article/n/obnaruzhenie-anomalnyh-sobytiy-na-hoste-s-ispolzovaniem-avtokodirovschika (дата обр. 21.02.2025).