Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.896

Басак
Дмитрий Владимирович

Модели и алгоритмы анализа, классификации и систематизации цифровых данных

АВТОРЕФЕРАТ
на соискание степени магистра
по специальности 1-40 80 01 «Компьютерная инженерия (встраиваемые системы)»

_____
(подпись магистранта)

Научный руководитель
Нестеренков Сергей Николаевич
(фамилия, имя, отчество)
к. т. н., доцент
(ученая степень, ученое звание)

_____
(подпись научного руководителя)

Минск 2024

ВВЕДЕНИЕ

Задача обнаружения аномалий в файлах журналов является широко изученной областью. Причина этого в том, что обнаружение аномалии может значительно помочь в обеспечении безопасности системы или компании. Файлы журнала содержат информацию о запуске системы и многих испектах ее работы. Возможность узнать, когда что-то идет не так, просто основываясь на информации лог-файла, очень ценна. Из-за этого было разработано бесчисленное множество подходов, включая подходы, использующие машинное обучение и, глубокое обучение.

Основной недостаток существующих подходов заключается в том, что они фокусируются только на конкретной проблеме или типе аномалии. В качестве альтернативы, некоторые существующие подходы адаптируются к конкретному набору данных и основаны на том, как этот набор может охарактиризовать эксперт с глубоким пониманием исходной системы. Оба этих подхода допустимы, но цель этой работы – предложить новое, более универсальное решение для обнаружения аномалий в файлах журналов, как с точки зрения аномалий, так и с точки зрения набора данных.

Для достижения основной цели будет проведен анализ существующих подходов, чтобы оценить, какие типы аномалий они способны обнаружить. Анализ будет проводиться как с размеченными общедоступными наборами данных, так и с наборами неразмеченных данных, предоставленных центром информатизации и инновационных разработок (ЦИИР). Оценка будет проводиться с учетом указанных требований со стороны подразделения ЦИИР, которое проявило интерес к использованию готового инструмента.

На основе оценки существующих подходов будет выявлено несколько возможных типов проблем, связанных с обнаружением аномалий в лог-файлах. Затем будет реализован программный прототип, которой сможет обнаруживать выбранные типы аномалий. Прототип будет реализован таким образом, что его можно будет использовать с несколькими наборами данных, так как будет предусмотрена настройка пользователем. Наконец, прототип будет оценен при решении задач определенных типов.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Файлы журналов могут содержать огромные объемы информации, особенно в крупных компьютерных системах, сетях и облачных сервисах. Обработка и анализ таких объемов данных вручную становится невыполнимой задачей, поэтому для обнаружения аномалий требуется

автоматизированный подход. Файлы журналов могут содержать разнообразную информацию, включая сообщения об ошибках, системные события, запросы к серверу и многое другое. Структура и формат этих данных могут быть сложными и разнообразными, что усложняет процесс анализа. Обнаружение аномалий в файлах журналов является важным инструментом обеспечения безопасности и надежности компьютерных систем. Аномалии могут указывать на попытки вторжения, атаки или сбои в работе системы, которые необходимо обнаружить и устранить.

В целом, поиск аномалий в лог-файлах является важным аспектом обеспечения безопасности, надежности и эффективности компьютерных систем и сетей, что делает эту проблему актуальной и востребованной в современном мире информационных технологий.

Данная диссертация также служит образовательным и исследовательским целям, позволяя получить представление о подходах к поиску аномалий в лог-файлах.

Степень разработанности проблемы

Исследование методов машинного обучения в области классификации записей в лог-файлах осуществлялись в подавляющем большинстве зарубежными авторами: Вэй Сюй, Лин Хуан, Фэй Тони Лю, Вэннел Зефак, Донхен Ким.

Возможность узнать, когда что-то идет не так, просто основываясь на информации лог-файла, очень ценна. Из-за этого было разработано бесчисленное множество подходов, включая подходы, использующие машинное обучение и глубокое обучение. Основной недостаток существующих подходов заключается в том, что они фокусируются только на конкретной проблеме или типе аномалии.

Цель и задачи исследования

Целью диссертации является разработка моделей и алгоритмов, позволяющих классифицировать записи в лог-файлах по признаку аномальности.

Поставленная цель работы определяет **следующие основные задачи:**

1. Провести обзор существующих методов поиска аномалий в файлах журналов.

2. Выбрать лучший метод каждого типа и провести сравнение на разных наборах данных.

3. Составить требования к обнаружению аномалий, классифицировать типы аномалий.

4. Разработать модели и алгоритмы для обнаружения аномалий с учетом установленных требований.

5. Эксперементально доказать эффективность разработанных моделей и алгоритмов.

Область исследования

Содержание диссертации соответствует образовательному стандарту высшего образования второй ступени (магистратуры) ОСВО 1-40 80 01-2019 специальности 1-40 80 01 Компьютерная инженерия (профилизация Встраиваемые системы).

Теоретическая и методологическая основа исследования

В основу диссертации легли работы зарубежных ученых в области поиска аномалий в файлах журналов различных компьютерных систем.

Информационная база исследования сформирована на основе литературы, открытой информации, технических нормативно-правовых актов, сведений из электронных ресурсов, а также материалов научных конференций и семинаров.

Научная новизна, теоретическая и практическая значимость

Научная новизна и значимость полученных результатов работы заключается в моделей и алгоритмов, позволяющих классифицировать записи в лог-файлах по признаку аномальности.

Теоретическая значимость работы заключается в подробном обзоре существующих подходов к обнаружению аномалий в лог-файлах.

Практическая значимость диссертации состоит в разработке моделей и алгоритмов для классификации данных в лог-файлах, что позволит упростить анализ, быстрее расследовать инциденты и увеличить надежность информационных систем.

Основные положения, выносимые на защиту

1. Систематизация методов поиска аномалий в файлах журналов, что позволяет определиться с наиболее подходящими для поставленных задач моделями.

2. Модели взаимодействия между функциональными компонентами программы в процессе парсинга лог-файлов, а также в процессе обучения и класификации данных.

3. Консольная утилита, включающая в себя несколько раширяемых компонентов, для поиска различных аномалий.

Апробация диссертации и информация об использовании ее результатов

Результаты исследований, вошедшие в диссертацию, опубликованы в сборниках материалов 59-ой научно-технической конференции аспирантов, магистрантов и студентов БГУИР (г. Минск, Беларусь, 2023 год), 9-ой международной научно-практической конференции *BIG DATA and Advanced Analytics* (г. Минск, Беларусь, 2023 год), международной научной конференции ИТС 2022 (г. Минск, Беларусь, 2022 год).

Публикации

Изложенные в диссертации основные положения и выводы опубликованы в 4 печатных работах, все из которых являются статьи в сборниках материалов научных конференций.

Общий объем публикаций по теме диссертации составляет 12 страниц.

Структура и объем работы

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, библиографического списка и приложений.

В первой главе представлен обзор методов для анализа лог-файлов и пойска аномалий, основы машинного обучения и теории графов.

Во второй главе представлено обоснование выбора определенных подходов в классификации данных в лог-файлах, дано описание используемых в работе наборов данных, а также произведен анализ результатов классификации данных с использованием выбранных подходов.

В третьей главе представлена архитектура утилиты для классификации данных, рассмотрены алгоритмы работы его компонентов. Описаны условия тестирования утилиты на наборах данных, а также анализ результатов тестирования.

В приложении представлены публикации автора и листинг кода.

Общий объем диссертационной работы составляет 91 страницу. Из них 63 страницы основного текста, 21 иллюстрации на 20 страницах, 8 таблиц на 6 страницах, библиографический список из 39 наименований на 3 страницах, список собственных публикаций соискателя из 4 наименований на 1 странице, 3 приложения на 20 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе вводятся и определяются все необходимые термины, которые используются в данной работе. Прежде всего, обсуждаются основные

понятия обнаружения аномалий. Обнаружение аномалий можно рассматривать как задачу классификации, имеющую два класса: нормальный и аномальный. Для измерения успешности классификации существует множество показателей. В первой главе описываются параметры точность, полнота, F1-мера, расмматриваются ситуации, в которых лучше использовать тот или иной параметр. Также объясняются основы машинного обучения и теории графов. Рассматриваются различные существующие подходы к поиску аномалий, такие как генеративные модели, рекуррентные нейронные сети, трансформеры, методы обработки естественного языка (Natural Language Processing, NLP), и многие другие.

Во второй главе предоставляется более подробный обзор выбранных подходов к обнаружению аномалий в журналах. Выбранные подходы охватывают несколько типов алгоритмов, включая стандартное машинное обучение и методы глубокого обучения. Первый важный шаг для многих методов обнаружения аномалий в лог-файлах – это так называемый синтаксический анализ журнала. Синтаксический анализ журнала используется для преобразования необработанных строк лог-файлов в структурированное представление. Следовательно, основной целью анализа журнала является идентификация и разделение всех шаблонов сообщений, которые встречаются в файле журнала. В частности, анализаторы удаляют переменные части, такие как имена пользователей, IP-адреса и другие идентификаторы, и оставляют шаблон журнала.

Также в этой главе рассматриваются технические детали выбранных подходов, а именно PCA [2], изоляционный лес [3] и LogCluster [21], представляющих стандартные подходы к машинному обучению. Кроме того, в этом разделе рассматриваются DeepLog [4], AutoLog [6] и LogBERT [5], основанные на глубоком обучении. Приводится описание наборов данных, которые будут использоваться в работе. Приводится сравнение подходов на уже размеченных наборах данных, результаты отраены в таблицах 1 и 2, а также на основе данных ЦИИР, в табице 3.

Таблица 1 – Результаты сравнения по набору данных HDFS

Подход	Полнота	Точность	F1-мера
RCA	0,673	0,998	0,804
Is. Forest	0,888	0,528	0,662
LogCluster	0,655	1,000	0,792
DeepLog	0,209	0,459	0,287
AutoLog	1,000	0,557	0,715
LogBert	0,178	0,428	0,257

Таблица 2 – Результаты сравнения по набору данных BGL

Подход	Полнота	Точность	F1-мера
RCA	0,233	0,114	0,153
Is. Forest	0,883	0,473	0,616
LogCluster	0,868	1,000	0,929
DeepLog	0,669	0,834	0,743
AutoLog	0,981	0,813	0,889
LogBert	0,663	0,923	0,772

Таблица 3 – Результаты анализа логов Linux

Подход	Обуч., с	Распозн., с	Всего	Ошибки	Трассир. стека	Без аномалий
RCA	74,93	43,63	10480	2665	1813	6055
Is. Forest	68,07	84,58	259	184	0	75
LogCluster	3499,68	1109,02	991	82	504	419
DeepLog	1653,00	495,62	141550	48313	1830	91574
AutoLog	593,96	39,94	3853	64	0	3789
LogBert	36718,59	312,84	3754	1127	19	2608

В третьей главе представлена архитектура утилиты для классификации данных, рассмотрены алгоритмы работы его компонентов. Описаны условия тестирования утилиты на наборах данных, а также анализ результатов тестирования. Общую архитектуру решения можно увидеть на рисунке 1. Решение предполагает использование необработанного файла журнала или файлов в формате csv. Затем необработанный журнал преобразуется в структурированное представление с извлечением переменных частей компонентом `ParserWrapper`.

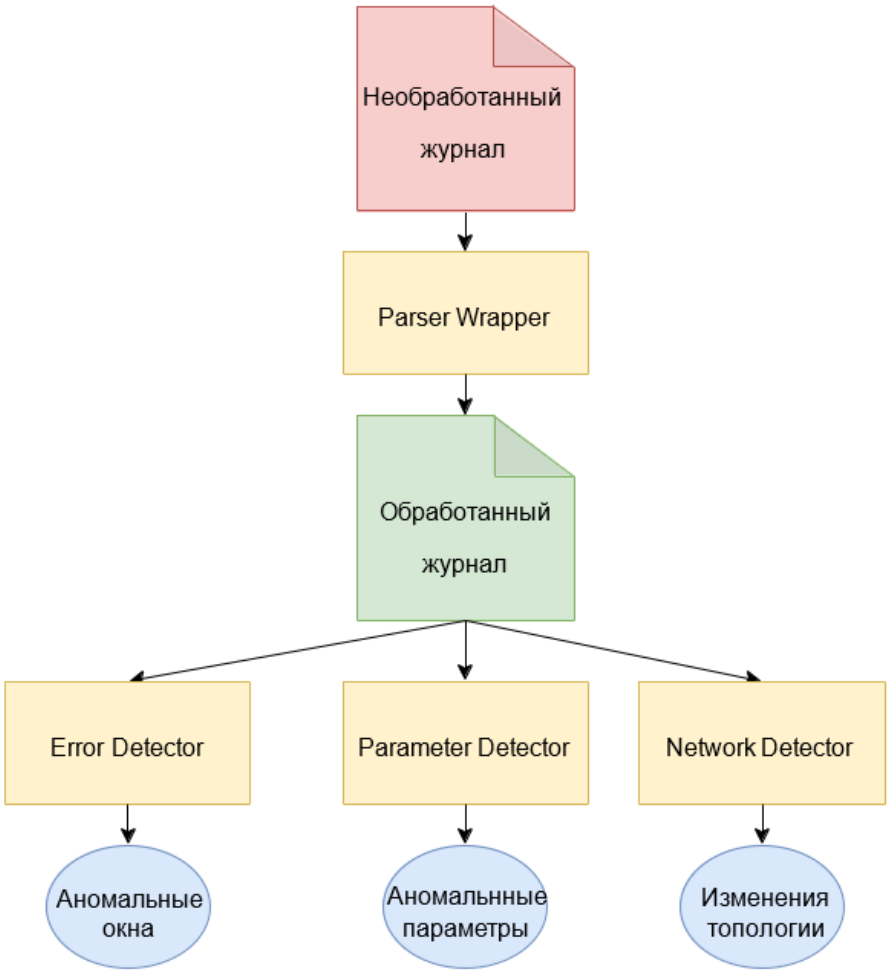

Рисунок 1 – Схема архитектуры решения

Проанализированное представление журнала затем используется модулями обнаружения. Эти модули обнаруживают различные типы аномалий. Существующие подходы могут быть применены для обнаружения ошибок, их причин и аномалий, проявляющихся в аномальном количестве сообщений журнала одного типа. Компонент `ErrorDetector` обнаруживает аномалии такого типа. Аномалии, которые могут характеризоваться аномально высокой частотой встречаемости лог-параметров, обнаруживаются компонентом `ParameterDetector`. Компонет `NetworkDetector` извлекает информацию о топологии сети из проанализированных журналов и затем использует эту информацию для обнаружения изменений топологии сети.

Затем была выполнена оценка, результаты которой приведены в таблице 4. Оценка была выполнена для каждого настроенного шаблона отдельно и для одновременного анализа всех шаблонов. В таблице результатов показано время прогнозирования, количество обнаруженных аномальных окон и процент правильно классифицированных окон. Оценка показала, что все окна, о которых сообщалось как об аномальных, были действительно аномальными

– это объясняется простой и понятной реализацией, основанной на настраиваемом пороге аномалий.

Таблица 4 – Результаты оценки детектора параметров

ID шаблона	Время, с	Количество окон	Найдено правильно, %
abb8e480	2,26	124	100
30661043	2,28	1488	100
78fa1af2	1,77	803	100
Все	2,96	2415	100

ЗАКЛЮЧЕНИЕ

Основной целью данной работы было предложить модели и алгоритмы классификации данных в лог-файлах по признаку аномальности. Основной вклад данной работы заключается в универсальности предлагаемого подхода. По сравнению с существующими подходами, предлагаемое решение было разработано путем предварительной оценки и классификации типов аномалий, а затем предложения и внедрения соответствующего решения для каждого из них. Более того, благодаря предлагаемым улучшениям, решение также является универсальным с точки зрения применимости к различным наборам данных.

Было проведено исчерпывающее экспериментальное сравнение выбранных подходов. Был проведен общий обзор стандартных методов машинного обучения и глубокого обучения, за которым последовало подробное описание и сравнение избранных подходов. Затем была проведена экспериментальная оценка подходов на соответствующих промаркированных наборах данных и сравнение на основе F1-меры. Подходы также были оценены на основе комплексного набора реальных данных, предоставленных ЦИИР. Цель этого сравнения состояла в том, чтобы оценить, какие аномалии могут быть обнаружены различными подходами с учетом установленных требований.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

Статьи в сборниках научных трудов

[1-А] Басак, Д. В. Проектирование личного кабинета сотрудника университета / Басак Д. В., Низовцов Д. В. // Электронные системы и технологии : сборник материалов 59-й научной конференции аспирантов,

магистрантов и студентов БГУИР, Минск, 17–21 апреля 2023 г. – Минск:БГУИР, 2023. – С. 68–70.

[2-А] Басак, Д. В. Информационная система управления студенческим общежитием / Басак Д. В., Низовцев Д. В., Нестеренков С. Н. // Информационные технологии и системы 2022 (ИТС 2022): материалы Международной научной конференции, Минск, 23 ноября 2022 – Минск : БГУИР, 2022. – С. 127–128.

[3-А] Нестеренков, С. Н. Применение Apache Spark для обработки больших данных в мобильных системах / С. Н. Нестеренков, Д. В. Басак, В. В. Куц // BIG DATA и анализ высокого уровня: сборник научных статей IX Международной научно-практической конференции, Минск, 17–18 мая 2023 г. : в 2 ч. Ч. 1 – Минск:БГУИР, 2023. – С. 379-381.

[4-А] Голубович, Ю. И. Применение больших данных в работе морских портов и терминалов / Ю. И. Голубович, С. Н. Нестеренков, Д. В. Басак // BIG DATA и анализ высокого уровня: сборник научных статей IX Международной научно-практической конференции, Минск, 17–18 мая 2023 г. : в 2 ч. Ч. 1 – Минск:БГУИР, 2023. – С. 272-275.