

**МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ им. Р.Е. АЛЕКСЕЕВА»**

ВЫПУСНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Ефодее Ирина Михайловна

Институт радиоэлектроники и информационных технологий

Кафедра «Вычислительные системы и технологии»

Группа М18-ИВТ-3

Дата защиты «09» июля 2020г.

Индекс
09.04.01

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ им. Р.Е. АЛЕКСЕЕВА»
(НГТУ)

Институт радиоэлектроники и информационных технологий
Направление подготовки (специальность) 09.04.01 «Информатика и вычислительная техника»
Направление (профиль) образовательной программы «Теоретическая информатика»
Кафедра «Вычислительные системы и технологии»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

магистра
(бакалавра, магистра, специалиста)

Студента Ефодее Ирины Михайловны Группы М18-ИВТ-3
На тему «Модель и алгоритмы идентификации пользователя по сетевому трафику»

СТУДЕНТ

И.Ефодее Ефодее И.М.
(подпись) (фамилия, и., о.)
02 июля 2020г.
(дата)

РУКОВОДИТЕЛЬ

В.Гай Гай В.Е.
(подпись) (фамилия, и., о.)
02 июля 2020г.
(дата)

РЕЦЕНЗЕНТ

М.Трескин Трескин М.И.
(подпись) (фамилия, и., о.)
02 июля 2020г.
(дата)

ЗАВЕДУЮЩИЙ КАФЕДРОЙ

Д.В.Жевнерчук Жевнерчук Д.В.
(подпись) (фамилия, и., о.)
02 июля 2020г.
(дата)

КОНСУЛЬТАНТЫ:

1. По _____

(подпись) (фамилия, и., о.)

(дата)

2. По _____

(подпись) (фамилия, и., о.)

(дата)

ВКР защищена «09» июля 2020г.
(дата)

Протокол № _____

С оценкой _____

Оглавление

1.	Обзор существующих методов	12
1.1.	Описание стека протоколов TCP/IP	12
1.2.	Обзор существующих методов решения задачи классификации трафика	13
1.2.1.	Классификация на основе блоков данных	14
1.2.2.	Классификация на основе статистического методов	15
1.3.	Структура алгоритмов машинного обучения	16
1.3.1.	Описание общего подхода к решению задач с использованием алгоритмов машинного обучения	17
1.3.2.	Обзор наиболее популярных алгоритмов машинного обучения	19
1.3.3.	Обзор метрик, определяющих адекватность модели	28
2.	Информационная модель классификации сетевого трафика	32
2.1.	Этап сбора данных для алгоритмов машинного обучения	33
2.2.	Выбор средства сбора сетевого трафика	34
2.3.	Сбор данных	38
2.3.1.	Описание сниффера WireShark	38
2.3.2.	Сбор данных в приложении WireShark	39
2.3.3.	Формирование признаков описания данных	42
2.4.	Выбор алгоритмов машинного обучения для решения поставленной задачи	44
2.5.	Выбор метрик оценки адекватности моделей	45
3.	Вычислительный эксперимент	46
3.1.	Описание эксперимента	46
3.2.	Описание программного продукта	47
3.3.	Подготовка данных для алгоритмов машинного обучения	49
3.4.	Постановка вычислительного эксперимента	51
3.5.	Сравнение результатов работы разных алгоритмов	55
3.6.	Сравнение результатов работы с другими исследованиями	57

ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)

Модель и алгоритмы
идентификации
пользователя по сетевому
трафику

Лит	Лист	Листов
	2	64
М18-ИВТ-3		

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ им. Р.Е. АЛЕКСЕЕВА»
(НГТУ)

АННОТАЦИЯ

к выпускной квалификационной работе

по направлению подготовки (специальности) 09.04.01 «Информатика и вычислительная техника»

студента Ефодее Ирины Михайловны группы М18-ИВТ-3

по теме «Модель и алгоритмы идентификации личности пользователя компьютера по сетевому трафику»

Выпускная квалификационная работа выполнена на 64 страницах, содержит 30 рисунков, 4 таблицы, 16 формул, библиографический список из 19 источников.

Актуальность:

Данный подход к идентификацию личности пользователя может быть использован для решения целого кластера задач, таких как повышения уровня защиты данных пользователя, улучшения качества работы алгоритмов контекстной рекламы, алгоритмов выборки актуальных новостей и т.д.

Объект исследования:

Объектом исследования являются сетевой трафик, сериализованный в один из общедоступных форматов хранения данных.

Предмет исследования:

Предметом исследования являются модели и методы решения задачи классификации сетевого трафика с использованием алгоритмов машинного обучения.

Цель исследования:

Разработка и исследование различных моделей и алгоритмов решения задачи идентификации пользователя компьютера по сетевому трафику с использованием признакового описания и алгоритмов машинного обучения.

Задачи исследования:

обзор и анализ существующих известных методов решения задачи классификации сетевого трафика с использованием признакового описания; создание информационной модели описания объекта с использованием признакового описания; создание алгоритма формирования признакового описания сетевого трафика; проведение исследования с целью

выявления наилучшей комбинации параметров и алгоритмов разрабатываемой системы; проведение вычислительного эксперимента для установления корректности работы созданных моделей и алгоритмов, а также оценка адекватности разработанных моделей

Методы исследования:

Методы формирования признаков описания сетевого трафика, наиболее популярные методы машинного обучения, метод вычислительного эксперимента, сбор данных с использованием анализатора сетевого трафика.

Структура работы:

3 раздела, введение, заключение и библиографический список.

Во введении отражены актуальность выбранной темы, цель работы и задачи исследования, научная новизна, теоретическая и практическая ценность работы, а также ее обоснованность и достоверность.

В разделе 1 «Обзор существующих методов» составлен обзор известных методов классификации сетевого трафика и алгоритмов машинного обучения, выявлены этапы решения этой задачи, а также проблемные места существующих методов.

В разделе 2 «Информационная модель классификации сетевого трафика» рассмотрены теоретические подходы к решению задачи классификации сетевого трафика на всех ее этапах, предлагаемые разработанным методом.

В разделе 3 «Вычислительный эксперимент» приведено описание вычислительного эксперимента, предназначенного для тестирования предлагаемого метода решения задачи, а также анализ результатов этого эксперимента.

В заключении обобщены результаты проделанной работы, сделаны выводы о достижении поставленной перед началом работы цели.

Выводы:

1. Разработанный метод идентификации пользователя компьютера по сетевому трафику дает корректные результаты работы, является конкурентоспособным по сравнению с аналогами, может использоваться на практике.
2. Задачи, поставленные перед началом исследования, выполнены, цель работы достигнута.

Рекомендации:

1. Рекомендуется использование результатов работы при формировании признаков описания сетевого трафика.
2. Рекомендуется использование результатов работы при создании систем идентификации пользователя по сетевому трафику и/или решению задачи классификации сетевого трафика.


(подпись)

/ Ефодее И.М.
(расшифровка подписи)

«02» июля 2020г.

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ им. Р.Е. АЛЕКСЕЕВА»
(НГТУ)

Кафедра «Вычислительные системы и технологии»


УТВЕРЖДАЮ
Заведующий кафедрой
«Вычислительные системы и технологии»
 **Жевнерчук Д.В.**
«02» июля 2020г.

ГРАФИК ПОДГОТОВКИ И ОФОРМЛЕНИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ



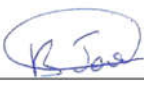
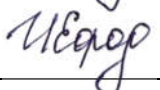
Студент:




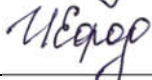
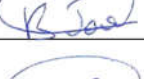
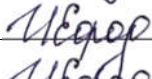
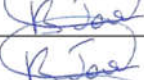
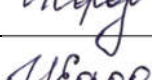

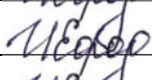
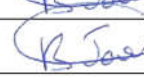
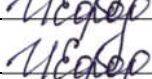




Ф.И.О.: Ефопе Ирина Михайловна
Группа: М18-ИВТ-3

Руководитель:

Ф.И.О.: Гай Василий Евгеньевич
Должность: доцент кафедры «ВСТ»
Ученое звание: доцент
Ученое звание: кандидат технических наук


Тема работы: «Модель и алгоритмы идентификации пользователя по сетевому трафику»

№	Этапы работы	Срок выполнения	Отметка о выполнении	
			Замечания руководителя	Подпись обучающегося
1	Подбор материала по теме ВКР, его изучение и обработка	06.04.2020		
2.	Разработка и представление руководителю первой части работы	20.04.2020		

3.	Разработка и представление руководителю второй части работы	06.05.2020		
4.	Разработка и представление руководителю третьей части работы	14.05.2020		
5.	Согласование ВКР с консультантами	14.05.2020		
6.	Подготовка и согласование с руководителем выводов и предложений	07.06.2020		
7.	Проверка нормоконтролера	23.06.2020		
8.	Получение отзыва руководителя ВКР	25.06.2020		
9.	Получение рецензии	26.06.2020		
10.	Представление ВКР заведующему кафедрой	02.07.2020		

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ им. Р.Е. АЛЕКСЕЕВА»
(НГТУ)

Кафедра _____ «Вычислительные системы и технологии» _____

УТВЕРЖДАЮ
Зав. Кафедрой «ВСТ»
 Жевнерчук Д.В.
«14» апреля 2020г.

ЗАДАНИЕ



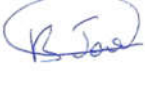


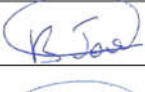
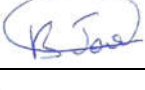
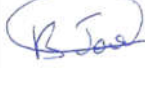
на выполнение выпускной квалификационной работы




по направлению подготовки (специальности) 09.04.01 «Информатика и вычислительная техника»

студенту Ефодее Ирине Михайловне группы М18-ИВТ-3

1. Тема ВКР «Модель и алгоритмы идентификации личности пользователя компьютера по сетевому трафику» (утверждена приказом по ВУЗу от 07.04.2020 № 845/5)
2. Срок сдачи студентом законченной работы: «02» июля 2020г.
3. Исходные данные к работе: Сериализованные в JSON данные, собранные с помощью анализаторов сетевого трафика Wireshark;
4. Содержание расчетно-пояснительной записки (перечень вопросов, подлежащих разработке):
 - Введение
 - 1. Обзор существующих методов
 - 2. Информационная модель классификации сетевого трафика
 - 3. Вычислительный эксперимент
 - Заключение
 - Библиографический список
5. Перечень графического материала (с точным указанием обязательных чертежей)
Общий объем работы – 61 страница. Содержит 30 рисунков, 4 таблицы, 16 формул.
Список литературы включает в себя 20 наименований.
6. Консультанты по ВКР (с указанием относящихся к ним разделов) _____
7. Дата выдачи задания 03.02.2020г.

Код и содержание компетенции	Задание	Проектируемый результат	Отметка о выполнении
ОК-1 способность совершенствовать и развивать свой интеллектуальный и общекультурный уровень	Выполнить исследование по теме распознавания эмоционального состояния диктора по голосу	Результаты проведенных исследований	
ОК- 2 способность понимать роль науки в развитии цивилизации, соотношение науки и техники, иметь представление о связанных с ними современных социальных и этических проблемах, понимать ценность научной рациональности и ее исторических типов	Иметь представление о роли науки в современном обществе и научной рациональности	Понимание ценности научной рациональности при проведении исследования	
ОК- 3 способность к самостоятельному обучению новым методам исследования, к изменению научного и научно-производственного профиля своей профессиональной деятельности	Изучение методов исследования в области распознавания	Знание методов исследования в области распознавания	
ОК-4, способность заниматься научными исследованиями	Выполнить исследование на тему «Модель и алгоритмы идентификации личности пользователя компьютера по сетевому трафику»	Результаты проведенных исследований, текст ВКР, научная новизна	
ОК- 5 использование на практике умений и навыков в организации исследовательских и проектных работ, в управлении коллективом	Изучение принципов организации командной работы в ходе разработки программного обеспечения (производственная и преддипломная практики)	Понимание принципов организации командной работы в ходе разработки программного обеспечения	
ОК- 6 способность проявлять инициативу, в том числе в ситуациях риска, брать на себя всю полноту ответственности	Изучение рисков и ответственности при разработке продукта	Понимание рисков и ответственности при разработке программных продуктов	
ОК-7, способность самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности	Разработка методов решения задачи идентификации пользователя по сетевому трафику, использование информационных технологий (язык программирования Python) при выполнении ВКР	Новые модели и алгоритмы, используемые для решения поставленной задачи, использованные информационные технологии, текст ВКР	
ОК-8 способность к профессиональной эксплуатации современного оборудования и приборов (в соответствии с целями магистерской программы)	Освоить использование программно-аппаратных комплексов для целей классификации сетевого трафика и промышленной разработки ПО	Описание использования программно-аппаратных комплексов для целей классификации сетевого трафика	
ОК-9, умение оформлять отчеты о проведенной научно-исследовательской работе и подготавливать публикации по результатам исследования	Подготовка отчета по распределенной НИР. Оформление пояснительной записки и графических материалов по ВКР. Подготовка публикации для конференции ИСТ-2020	Текст отчета по распределенной НИР, текст пояснительной записки, текст публикации, графические материалы	
ОПК-1 способность воспринимать математические, естественнонаучные, социально-экономические и профессиональные знания, умение самостоятельно приобретать, развивать	Использовать математические, естественнонаучные, социально-экономические и	Использование математических, естественнонаучных, социально-	

и применять их для решения нестандартных задач, в том числе, в новой или незнакомой среде и в междисциплинарном контексте	профессиональные знания для решения задач классификации сетевого трафика и разработки ПО	экономических и профессиональных знания для решения задач классификации сетевого трафика и разработки ПО	
ОПК-2, обладать культурой мышления, способностью выстраивать логику рассуждений и высказываний, основанных на интерпретации данных, интегрированных из разных областей науки и техники, выносить суждения на основании неполных данных	Выполнение обзора и анализа методов идентификации пользователя по сетевому трафику, выбор инструментов и методов сбора и обработки данных предметной области, анализ и интерпретация данных предметной области, предложениивариантов применения разработанных моделей и методов в других областях	Пояснительная записка к ВКР, выступление на защите ВКР	
ОПК-3, обладать способностью анализировать и оценивать уровни своих компетенций в сочетании со способностью и готовностью к саморегулированию дальнейшего образования и профессиональной мобильности	Оценка результатов выполнения ВКР, анализ полученных результатов	Варианты дальнейшего развития исследования, отраженные в пояснительной записке	
ОПК-4, владением, по крайней мере, одним из иностранных языков на уровне социального и профессионального общения, способностью применять специальную лексику и профессиональную терминологию языка	Использование при выполнении ВКР литературы зарубежных авторов на английском языке	Список литературы с включенными в него зарубежными источниками	
ОПК-5 владение методами и средствами получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях	Использовать методы и средства получения, хранения, переработки и анализа информации в Internet, хранение исходного кода и результатов эксперимента в общем доступе	Использование методов и средств получения, хранения, переработки и анализа информации в Internet, хранение исходного кода и результатов эксперимента в общем доступе	
ОПК-6, обладать способностью анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациям	Составление обзора методов классификации сетевого трафика, структурирование полученной информации, составление текста выступления на защите ВКР	Обзор методов классификации сетевого трафика, текст пояснительной записки, текст и презентация выступления на защите ВКР	
ПК-1 знание основ философии и методологии науки	Знание основ методологии научных исследований	Применение методов научных исследований	
ПК-2 знание методов научных исследований и владение навыками их проведения	Изучение методов системного анализа при проведении исследования	Применение системного анализа при проведении исследования	
ПК-3, знанием методов оптимизации и умение применять их при решении задач профессиональной деятельности	Выполнить подбор входных параметров разработанного алгоритма, обеспечивающих наибольшую точность локализации и классификации	Результаты вычислительного эксперимента в пояснительной записке	

ПК-5 владение существующими методами и алгоритмами решения задач цифровой обработки сигналов	Изучение методов дискретизации сигналов во времени	Знание методов дискретизации сигналов во времени	
ПК-6 понимание существующих подходов к верификации моделей программного обеспечения (ПО)	Изучение методов верификации ПО	Метод верификации в пояснительной записке	
ПК-7, применением перспективных методов исследования и решения профессиональных задач на основе знания мировых тенденций развития вычислительной техники и информационных технологий	Использование алгоритмов машинного обучения, методов классификации сетевого трафика, языка программирования Python	Разработанные модели и алгоритмы, результаты вычислительного эксперимента, сравнение результатов между собой и с результатами аналогов	

Руководитель


(подпись)

В.Е. Гай

(И.О. Фамилия)

Задание к исполнению принял «03» февраля 2020г.
(дата)

Студент


(подпись)

И.М. Ефод

(И.О. Фамилия)

Введение

Актуальность проблемы

Сетевой трафик (в литературе широко используется понятие «интернет-трафик», но это понятие не совсем корректно, так как обмен информации может происходить без использования всемирной сети Интернет) представляет собой информацию, передаваемую посредством какой-либо компьютерной сети с использованием определенных правил (протоколов) и за определенное время.

Практически каждый сегодня владеет сразу несколькими устройствами (Например, смартфоны, планшетные компьютеры, ноутбуки, рабочие станции и т.п.), которые активно использует для обмена, получения и передачи информации в различных целях. Из этого можно сделать вывод, что конкретный пользователь генерирует уникальный сетевой трафик, который определяется поведенческими привычками пользователя и характеристиками сетевых сессий. Следовательно, между уже собранным сетевым трафиком и новыми данными, собранными за некоторый период времени, существует достаточно сильная корреляция. Это может позволить достаточно точно определить конкретного пользователя по данным использования сети.

Стоит отметить, что при недостаточном объеме данных для обучения алгоритма (и особенно, при использовании статистики только с одного из устройств), данный подход не только не покажет своей эффективности, но и с высокой долей вероятности, будет давать ложные предсказания.

Обладая достаточным объемом данных большого числа пользователей, предложенный подход позволит не только идентифицировать конкретного человека (из числа представивших свой трафик), но и предсказывать некоторые признаки, описывающие каждого конкретного пользователя (Например, пол, возраст и т.д.) или группу пользователей, что позволяет значительно улучшить работу алгоритмов контекстной рекламы, адресного подбора новостей, улучшение алгоритмов безопасного использования устройства с целью хранения данных и т.д. и т.п., то есть решать широкий спектр задач.

Цель работы и задачи исследования

Целью работы является исследование различных методов классификации сетевого трафика, разработка модели идентификации пользователя по сетевому трафику с использованием алгоритмов машинного обучения, разработка качественного признакового описания объектов и сравнение полученных результатов между собой и с аналогами.

Для достижения этой цели необходимо решить следующие задачи:

- Обзор и анализ известных методов решения задачи классификации сетевого трафика.
- Сбор данных для последующего обучения алгоритмов.
- Создание алгоритма формирования признакового описания трафика, то есть выделение признаков, описывающих не только характеристики сетевых сессий, но и поведенческие привычки конкретного пользователя.
- Создание информационной модели описания объектов для формирования обучающей и валидационной выборок.
- Проведение исследования с целью выявления наиболее эффективных алгоритмов для решения данной задачи.

Ине. № подл	Подп. и дата
Ине. № дубл.	Взам. инв. №
Подп. и дата	Подп. и дата
Ине. № подл	Подп. и дата

Ли	Изм.	№ докум.	Подп.	Дат

- Подборка параметров для разрабатываемой системы.
- Проведение вычислительного эксперимента для установления корректности работы реализованных алгоритмов.
- Создание программной системы распознавания личности пользователя компьютера на основе сетевого трафика.

Объект исследования

Объектом исследования являются сетевой трафик, сериализованный в один из общедоступных форматов хранения данных.

Предмет исследования

Предметом исследования являются модели и методы решения задачи классификации сетевого трафика с использованием алгоритмов машинного обучения.

Методы исследования

В процессе исследования были использованы методы:

- статистического анализа полученных данных;
- методы классического машинного обучения для выявления зависимостей;
- вычислительного эксперимента, результатом которого является программный продукт, разработанный с использованием языка программирования высокого уровня Python.

Положения, выносимые на защиту

На защиту выносятся следующие положения:

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	<div>ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)</div>	Лист	
							9
Ли	Изм.	№ докум.	Подп.	Дат			

- Модель идентификации пользователя по сгенерированному сетевому трафику;
- Статистические алгоритмы классификации сетевого трафика;
- Результаты проведенных вычислительных экспериментов и их сравнение.

Научная новизна

В ходе написания работы был предложен новый способ идентификации личности пользователя компьютера, заключающийся в статистическом анализе сетевого трафика на базе выделения поведенческих привычек пользователя и описания сетевых сессий с использованием алгоритмов классического машинного обучения. Главное отличие данного метода от аналогов (например, cookie-файлов) заключается в отсутствии каких-либо данных, хранящихся и выполняющихся на клиентской стороне, а также анализе поведенческих привычек пользователя на базе сетевого трафика и отсутствии общепринятых данных для идентификации пользователя (Пол, фото, отпечатки пальцев, голос и т.д.). Это позволяет идентифицировать пользователя полностью на стороне сервера, а так же решить довольно обширный спектр задач.

Теоретическая и практическая ценность

Теоретическая ценность работы заключается в разработке информационной модели системы распознавания, а также в предложенном алгоритме идентификации пользователя по сетевому трафику.

Практическая ценность работы заключается в созданном программном продукте, реализующем указанный выше алгоритм, а также в результатах проведенного вычислительного эксперимента.

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	Лист	10
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)	

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата

Ли	Изм.	№ докум.	Подп.	Дат

Объём и структура

И. М. Ефодее, В.Е. Гай МОДЕЛЬ И АЛГОРИТМЫ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЯ КОМПЬЮТЕРА ПО СЕТЕВОМУ ТРАФИКУ (Нижегородский государственный технический университет им. Р.Е. Алексеева, г. Нижний Новгород)

Выпускная квалификационная работа состоит из введения, трех глав основной части, заключения и списка литературы. Общий объем работы – 64 страниц. Диссертация содержит 30 рисунков, 4 таблицы, 16 формул. Список литературы включает в себя 20 наименований.

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата

Ли	Изм.	№ докум.	Подп.	Дат

ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)

Лист
12

- Прикладной (Application)

- п) - на данном слое работает большинство приложений и т.д.) (HTTP, HTTPS, DNS, FTP и т.д.)
- rt) - доставляют пакеты данных, гарантия правильного приема данных (TCP, UDP и т.д.)
- ная задача доставка данных от отправителя получателю. В сетях работают маршрутизаторы (IP)
- ccess) - кодирует данные на физическом уровне (Ethernet, WLAN и т.д.)
- с моделью OSI-ISO приведено на рис. 1.1
- ов пользуются почти все приложения, используемые в сетях. При этом системная информация передается другими способами.
- | | |
|--------------------------------------------|-------------------|
| ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ) | <i>Лист</i>
12 |
|--------------------------------------------|-------------------|

Данным стеком протокол

ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)	Лист 12
--------------------------------------------	------------

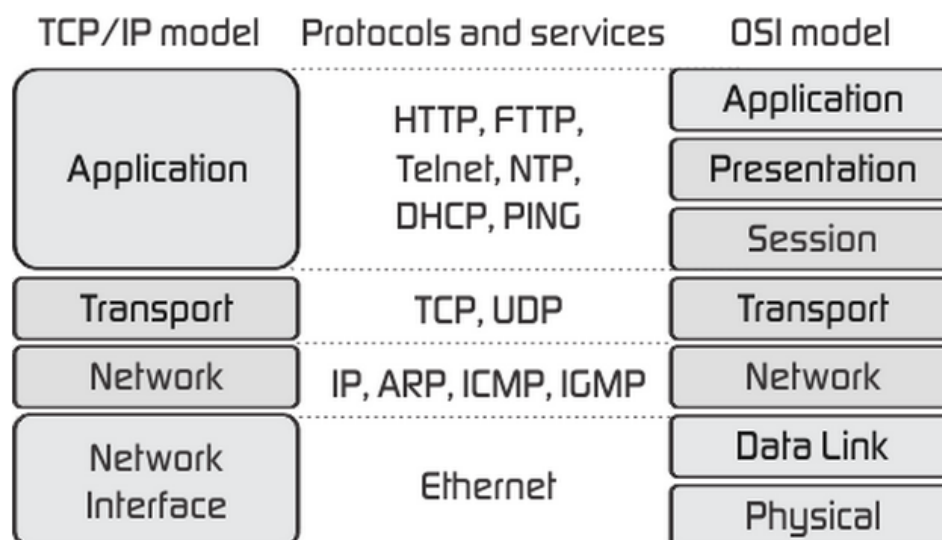


Рисунок 1.1. Сравнение стека TCP/IP с моделью OSI-ISO

ми способами. То есть именно статистика, собранная для этого стека протоколов может быть использована для решения задачи идентификации личности.

1.2. Обзор существующих методов решения задачи классификации трафика

Несмотря на то, что создано огромное множество различных алгоритмов для решения задачи классификации сетевого трафика, в настоящее время используется 2 подхода к решению данной задачи [2] (см. рис. 1.2)

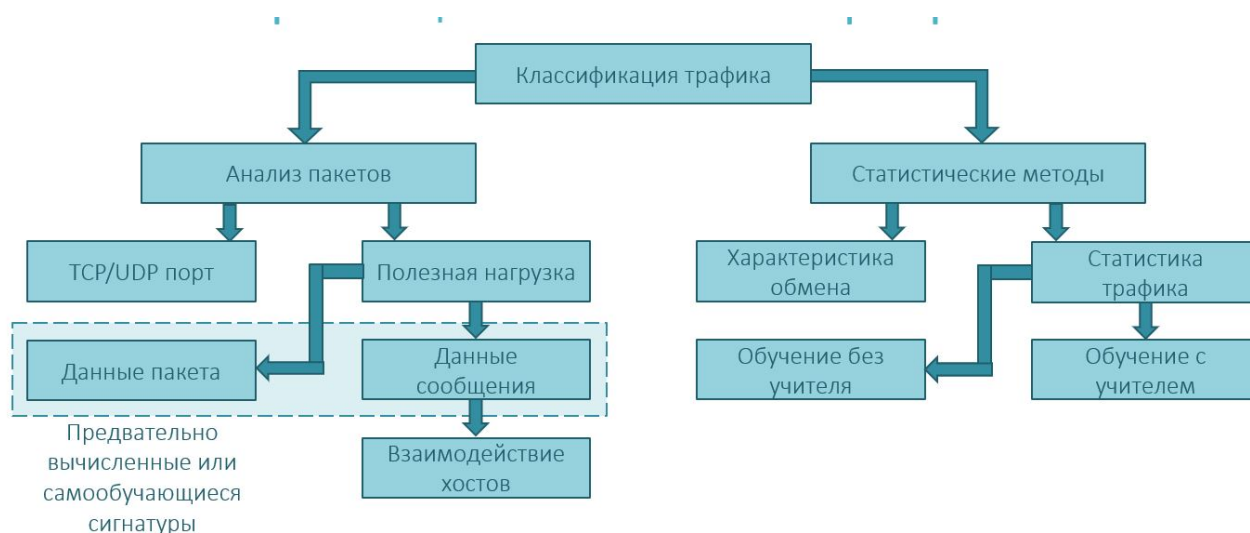


Рисунок 1.2. Основные подходы к решению задачи классификации сетевого трафика

Опишем эти подходы[3]:

- Классификация на основе блоков данных (Payload-Based Classification, Deep Package Inspection). Метод, основанных на полях с блоками информации, таких как ТСР-порты (отправитель, получатель или оба), а также содержимом самих пакетов. Наиболее распространенный метод для решения задачи классификации сетевого трафика, но не способен работать с зашифрованным и туннелированным трафиком, что означает привлечение дополнительных вычислительных ресурсов и специализированных алгоритмов для решения задачи для расшифровки информации.
- Классификация на основе статистического метода (Statistical Analysis). Данный подход основывается на статистическом анализе основных характеристик сетевых сессий (например, время и длительность сессии, местоположение пользователя, ip-адреса и т.д. и т.п.).

1.2.1. Классификация на основе блоков данных

При использовании этого метода выделяют универсальный подход к классификации сетевого трафика, а также способ глубокого анализа пакетов (Deep Package Inspection - DPI).

Универсальный подход к решению задачи классификации трафика основывается на данных в заголовке IP-пакетов – обычно это IP-адрес, MAC-адрес и используемый протокол [4]. Стоит отметить ограниченные возможности этого способа, поскольку информация берется только из IP-заголовка, так же, как ограничены методы с использованием портов – так как не все приложения используют стандартные порты, а кроме этого часто они могут быть обозначены пользователем.

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата						
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					Лист
										14

Более точный результат решения задачи классификации позволяет получить глубокий анализ пакетов (DPI). Такие системы позволяют распознавать приложения и протоколы, которые невозможно определить на сетевом уровне, (например, URL, содержимое сообщений мессенджеров, голосовой трафик Skype, р2р-пакеты BitTorrent и т.д. и т.п.). Из этого следует, что DPI анализирует не только заголовки, но и полное содержимое пакетов на всех уровнях модели ISO-OSI, начиная с канального уровня [5].

Основным механизмом идентификации приложений в DPI является анализ сигнатур (Signature Analysis). Все приложения имеют свои уникальные характеристики, занесенные в базу данных сигнатур. Сопоставление образца из базы данных с образом анализируемого трафика позволяет достаточно точно определить приложение или протокол.

Приложения регулярно обновляются и пополняются, что означает, что база данных сигнатур нуждается в регулярных обновлениях для обеспечения высокой точности алгоритмов работы [6].

Перечислим несколько методов сигнатурного анализа:

- Анализ образца (Pattern analysis)
- Числовой анализ (Numerical analysis)
- Поведенческий анализ (Behavioral analysis)
- Эвристический анализ (Heuristic analysis)
- Анализ протокола/состояния (Protocol/state analysis)

1.2.2. Классификация на основе статистического методов

В статистических методах необходимо различать два разных подхода:

- поведенческие алгоритмы и статистические алгоритмы сетевого уровня

Инв. № подл.	Подп. и дата					Лист 15
	Взам. инв. №					
	Инв. № дубл.					
	Подп. и дата					
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)	

- поведенческие алгоритмы и статистические алгоритмы транспортного уровня

Основная цель поведенческих алгоритмов метода состоит в определении приложений, создающих потоки сетевого трафика. Анализируя взаимодействие хостов в компьютерной сети, можно определить приложения, запущенные на компьютере. Отношения между классом трафика и его поведенческими статистическими свойствами были описаны в специализированных базах данных, включающих в себя эмпирические модели характеристик соединения для ряда специальных ТСП-приложений

Данный подход опирается на статистические характеристики трафика для идентификации приложения. Основа таких методов заключается в том, что сетевой трафик обладает статистическими характеристиками, являющимися уникальными для определенных классов приложений и позволяют разделить трафик по приложениям.

Статистические алгоритмы делятся на две группы [7]:

- методы классификации или обучение с учителем
- методы кластеризации или обучение без учителя

Главным недостатком такого подхода является неточность алгоритмов при неоднородном распределении данных.

1.3. Структура алгоритмов машинного обучения

Машинное обучение – класс методов из области искусственного интеллекта, изучающих алгоритмы, способные обучаться, то есть для которых характерно обучение в процессе применения решений множества похожих задач. В общем случае, задачи, которые решаются с использованием алгоритмов машинного обу-

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата						Лист
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					16

чения, сводятся к выявлению зависимостей (в том числе и скрытых) между некоторым набором признаков [8].

1.3.1. Описание общего подхода к решению задач с использованием алгоритмов машинного обучения

Для обучения алгоритмов машинного обучения нужны размеченные данные (Например, структурированные таблицы или базы данных), в которых содержится некоторое множество объектов (или ситуаций) и множество возможных ответов (или реакций). Главная задача алгоритмов данного класса – выявить любого рода корреляции между объектами и ответами, которые изначально не определены. Однако, существует конечная совокупность некоторого числа прецедентов (пара – «объект»-«ответ»), которую принято называть обучающей выборкой. Именно на этой основе применяемый алгоритм должен восстановить зависимости, что означает научиться давать точный классифицирующий ответ для любого возможного входного объекта. Выявляемая зависимость может быть выражена не только аналитически, но и эмпирически, то есть система должна обучиться обобщать любые возможные входные данные, в том числе и выходящие за пределы обучающей выборки. Для оценки качества работы полученной системы вводится оценка функционала качества.

Исходя из вышесказанного, можно сказать, что задача машинного обучения сводится к аппроксимации функций зависимости ответов системы от входных данных

На вышеуказанной диаграмме представлена последовательность решения задач с использованием алгоритмов машинного обучения. По сути, эффективность и корректность работы алгоритмов зависит от подобранных данных, выбора модели и параметров.

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата						
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					Лист
										17

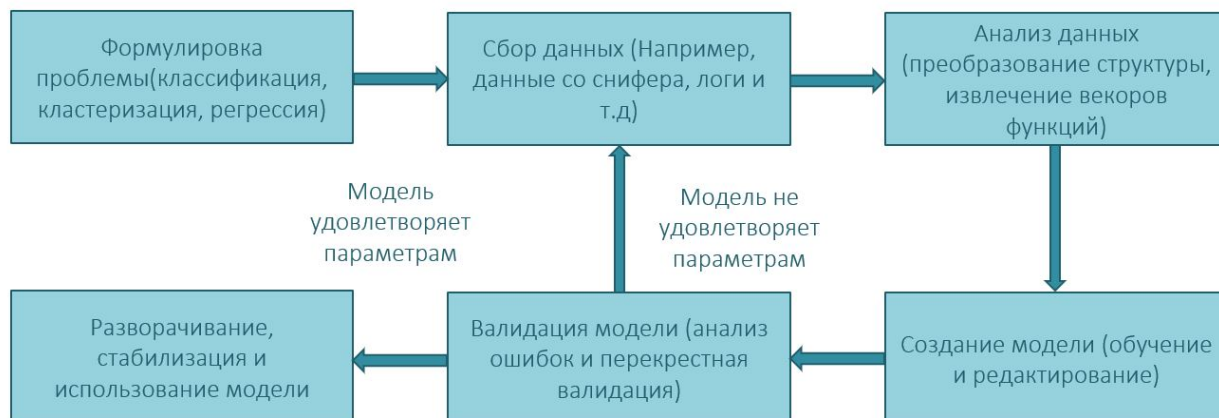


Рисунок 1.3. Основные этапы решения задачи классификации сетевого трафика

Следует отметить, что существует проблема переобучения модели, то есть система выдает хорошие результаты на тестовой выборке, но при этом очень плохо работает на примерах, не участвующих в обучении (на валидационных и реальных данных). Данная сложность возникает при использовании большого количества однообразных данных для обучения, а также слишком большом количестве эпох обучения. Кроме вышеуказанной, существует противоположная – проблема недообучения, проявляющаяся не только из-за неправильно подобранных данных, но и недостаточно сложных моделей. Во избежание данных сложностей при создании программного продукта, нужно регулярно сравнивать ошибки как на обучающей выборке, так и на валидационной выборках – они не должны сильно отличаться.

Задачу идентификации пользователя компьютера по сетевому трафику можно отнести к классу задач классификации, так как множество объектов разделены некоторым образом на классы. Количество объектов конечно, и заранее известно к каким классам они относятся (что касается обучающей выборки) [7].

Если говорить о конкретной ситуации, к классу мы отнесем конкретного человека, а к объектам статистику использования сетевого трафика этим пользователем.

Инва. № подл.	Подп. и дата	Взам. инв. №	Инва. № дубл.	Подп. и дата	Инва. № подл.
Ли	Изм.	№ докум.	Подп.	Дат	Лист
ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					18

1.3.2. Обзор наиболее популярных алгоритмов машинного обучения

Алгоритмы машинного обучения[9] представляют собой вычисление и обучение некоторой целевой функции - $f(x)$, которая наилучшим образом классифицирует входные переменные - X и выходную переменную - Y : $Y = f(X)$.

Так как неизвестно, что представляет собой функция f , ее нужно найти используя различные алгоритмы.

Одной из самых распространённых задач для алгоритмов машинного обучения является предсказание значений $Y = f(x)$ для новых входных значений - X , это называется прогностическим моделированием, цель которого – дать максимально точное и верное предсказание.

Разберем самые популярные алгоритмы машинного обучения[8]:

- Алгоритмы линейной регрессии.

Главной целью прогностического моделирования является минимизация ошибки модели, то есть более точное прогнозирование. Линейную регрессию можно описать уравнением, описывающее прямую, максимально точно показывающую взаимосвязь между входными данными X и выходными - Y . Главной задачей линейной регрессии является нахождение коэффициентов B для входных переменных, которые обеспечивают максимально точный результат.

Регрессионную модель можно представить как $y = f(x, b) + \epsilon$, $E(\epsilon) = 0$, где b - параметры модели, ϵ - случайная ошибка модели, а $f(x, b)$ имеет вид:

$$f(x, b) = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n,$$

где n - количество параметров модели.

Для оценки точности регрессионной модели используются методы линейной алгебры или статистические (например, метод наименьших квадратов).

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата						
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					Лист
										19

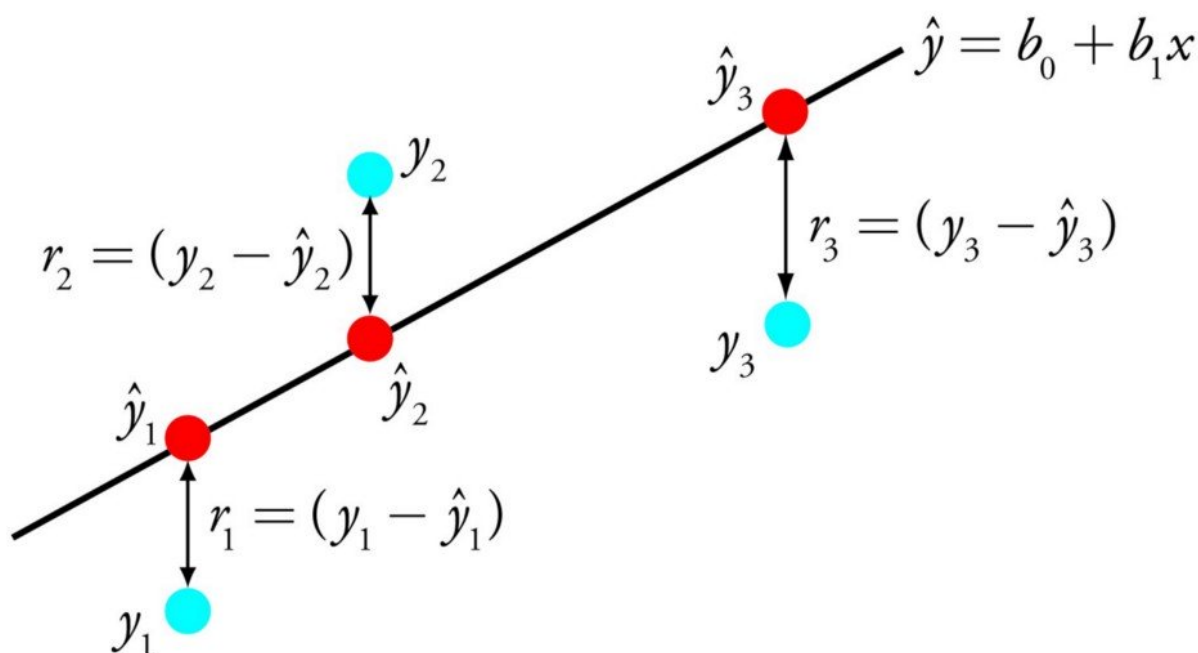


Рисунок 1.4. Иллюстрация работы алгоритма линейной регрессии

- Логистическая регрессия.

Логистическая регрессия является своего рода аналогом линейной, так как при этом подходе нужно найти значения коэффициентов для входных данных, однако значение выходной функции преобразуется с помощью нелинейной логистической функции (см. рис. 1.7).

Логистическая регрессия описывается формулой:

$$f(x) = \frac{1}{1+e^{-x}},$$

где x - входное значение.

Логистическая функция проводит преобразование любого числа в число от 0 до 1, что может быть использовано для предсказания класса в случае бинарной классификации.

- Линейный дискриминаторный анализ (LDA).

Инв. № подл	Подп. и дата				Лист 20
	Взам. инв. №				
	Инв. № дубл.				
	Подп. и дата				
ВКР-НГТУ-09.04.01-(М18-ИБТ-3)-006-2020(ПЗ)					Лист 20
Ли	Изм.	№ докум.	Подп.	Дат	

ных, однако значение выходной функции преобразуется с помощью нелинейной логистической функции (см. рис. 1.7).

Логистическая регрессия описывается формулой:

$$f(x) = \frac{1}{1+e^{-x}},$$

где x - входное значение.

Логистическая функция проводит преобразование любого числа в число от 0 до 1, что может быть использования для предсказания класса в случае бинарной классификации.

- Линейный дискриминаторный анализ (LDA).

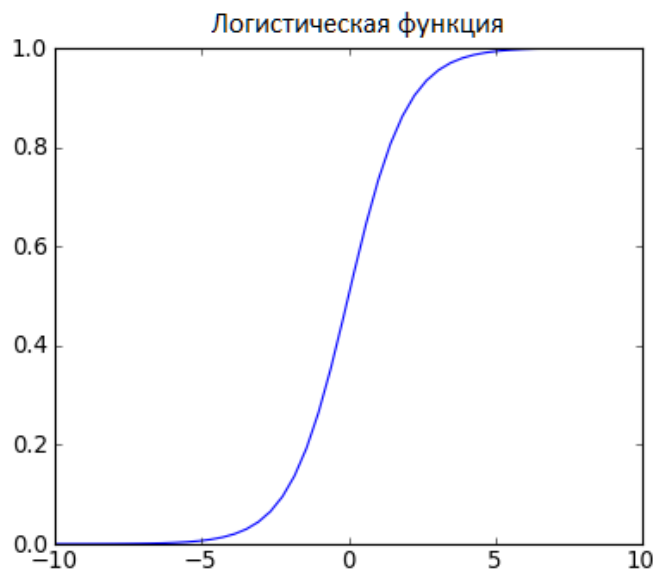


Рисунок 1.5. Логистическая функция

Логистическая регрессия позволяет определять принадлежность объекта к одному из классов. При большом количестве классов, рекомендуется использовать алгоритм LDA - Linear discriminant analysis.

LDA состоит из статистических свойств данных, рассчитанных для каждого класса:

- Поклассовое среднее значение;
- Поклассовая дисперсию.

Результат вычисляется с использованием поклассового дискриминантного значения с целью последующего выбора класса с максимальным результатом. Данные предполагают нормальное распределение, поэтому перед началом работы рекомендуется провести над данными некоторые преобразования, например, удалить аномальные значения, не подпадающие под нормальное распределение.

Описать данный классификатор можно функцией:

$$\sum_b = \frac{1}{C} \sum_{i=1}^C (\mu_i - \mu)(\mu_i - \mu)^T,$$

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата
Ли	Изм.	№ докум.	Подп.	Дат
ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)				
				Лист
				21

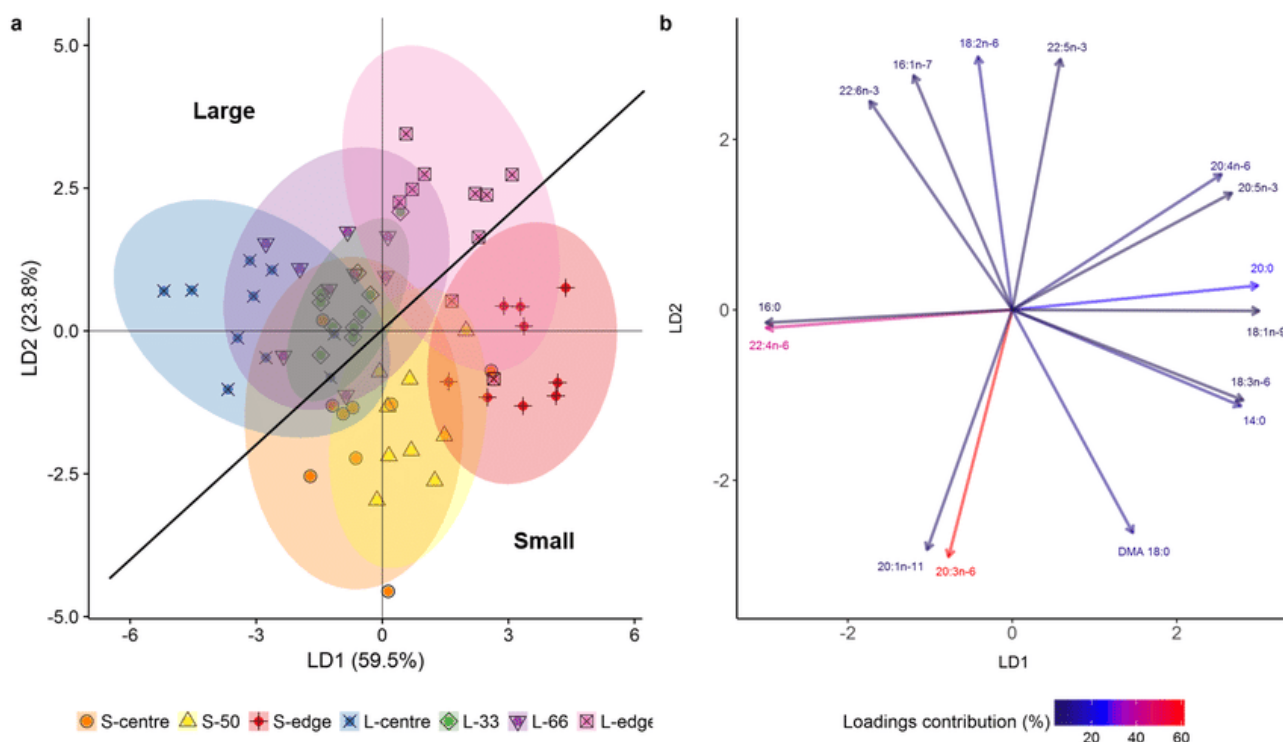


Рисунок 1.6. Иллюстрация работы алгоритма LDA

где μ - среднее средних для всех классов, C - количество классов.

Данный алгоритм требует нормализации данных.

- Деревья принятия решений.

Дерево решений чаще всего изображается в виде двоичного дерева, каждый узел которого - входная переменная, листовые узлы - выходная переменная, используемая для предсказания. Оно производится путём прохода по дереву к листовому узлу и вывода значения класса на этом узле.

Деревья быстро обучаются, и делают достаточно точные предсказания. Они предоставляют точные результаты для широкого круга задач и не требуют особой подготовки данных, однако могут занимать огромное количество памяти, так как нуждаются в постоянных обновлениях баз знаний.

- Наивный Байесовский классификатор.

Модель состоит из двух вероятностей, рассчитывающихся на основе тренировочных данных:

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)
					Лист 22

- Вероятность появления каждого класса
- Условная вероятность для каждого класса при каждом значении x .

После расчёта вероятностной модели, её можно использовать для работы с новыми данными при помощи теоремы Байеса:

$$p(C|F_1, \dots, F_n) = \frac{p(C)p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}$$

Вещественные данные предполагают их нормальное распределение.

- К-ближайших соседей (KNN).

Модель KNN (K-nearest neighbors) представлена всем набором тренировочных данных. Предсказание для новой точки делается путём поиска K ближайших соседей в наборе данных и суммирования выходной переменной для этих K экземпляров.

Для каждого класса j :

$$Q_j = \sum_{i=1}^n \frac{1}{d(x, a_i)^2}$$

,где $d(x, a_i)$ - расстояние от нового значения x до объекта a_i .

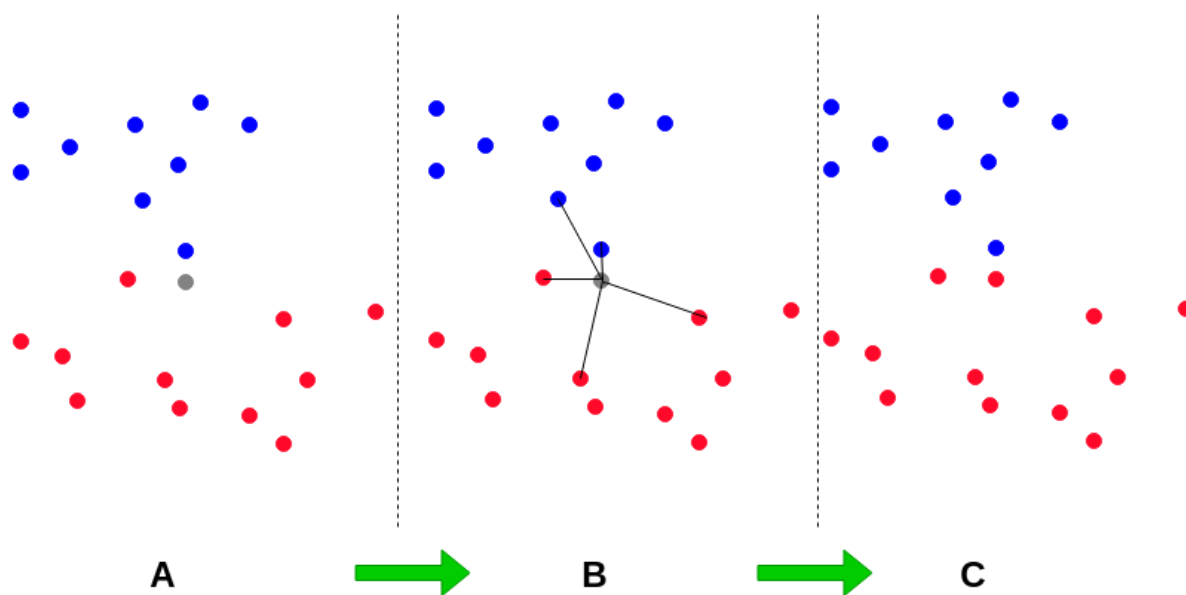


Рисунок 1.7. Иллюстрация работы KNN

Подп. и дата	
Взам. инв. №	
Инв. № дубл.	
Подп. и дата	
Инв. № подл.	

Ли	Изм.	№ докум.	Подп.	Дат

Самый простой способ определения принадлежности класса заключается в использовании евклидова расстояния — числа, которое можно рассчитать на основе различий с каждой входной переменной, однако все переменные должны иметь один масштаб.

KNN может требует большого количества памяти для хранения данных, но способен быстро и точно делать предсказание. Обучающую выборку можно регулярно обновлять для обеспечения высокой точности работы модели с течением времени.

Алгоритм ближайших соседей некачественно работает на многомерных данных (множество входных переменных), что негативно сказывается на эффективности алгоритма при решении задачи (чаще всего такое явление называют проклятием размерности). Поэтому для получения качественного результата, количество данных нужно минимизировать, то есть проводить некоторую дополнительную обработку данных, удаляя "ненужные" данные.

- Метод опорных векторов (Support Vector Machine - SVM).

Основная идея данного метода заключается в конверсии исходных векторов, описывающих входные данные, в более многомерное пространство и поиск в нем гиперплоскости, разделяющей классы, (пример 3.5) с максимальным зазором. Работа алгоритма основывается на предположении, что большая разница(расстояние) между этими параллельными гиперплоскостями обеспечивает меньшую среднюю ошибку классификатора.

Главная задача алгоритма сводится к поиску гиперплоскости, максимально точно описывающей разделение данных на классы. В процессе обучения алгоритм ищет коэффициенты функции, описывающей гиперплоскость.

Все данные описываются как набор точек:

Инв. № подл	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата	Инв. № подл	Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)	Лист
												24

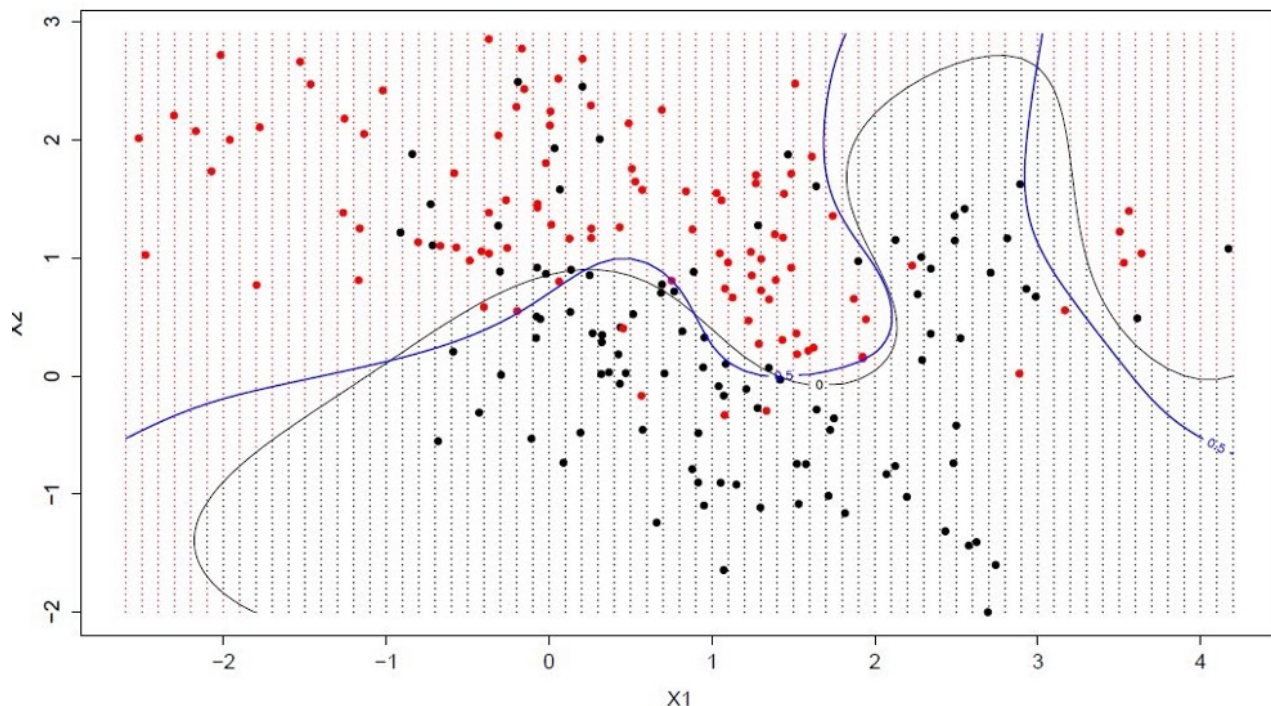


Рисунок 1.8. Иллюстрация работы SVM с данными, разделенными на 2 класса

$$(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \dots, (\mathbf{x}_n, c_n),$$

где c_i принимает значение 1 или -1 , в зависимости от того, какому классу принадлежит точка \mathbf{x}_i . Каждое \mathbf{x}_i — это p -мерный вещественный вектор, обычно нормализованный значениями $[0, 1]$ или $[-1, 1]$. Если данные не нормализованы, то выброс (точка с большими отклонениями от средних значений координат точек) может оказать сильное влияние на модель.

Гиперплоскость имеет вид:

$$wx - b = 0,$$

где w - перпендикуляр к разделяющей гиперплоскости. Параметр $\frac{b}{\|\mathbf{w}\|}$ равен модулю расстояния от гиперплоскости до начала координат.

Если классы линейно неразделимы, то алгоритм может допускать ошибки на обучающей выборке. $\xi_i \geq 0$ - набор переменных, характеризующих величину ошибки на объектах \mathbf{x}_i , $1 \leq i \leq n$. Введём в минимизируемую функцию штраф за суммарную ошибку:

$$\begin{cases} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \rightarrow \min_{w,b,\xi_i} \\ c_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, \quad 1 \leq i \leq n, \\ \xi_i \geq 0, \quad 1 \leq i \leq n \end{cases}$$

где коэффициент C — параметр настройки метода, позволяющий регулировать отношение между максимальной шириной разделяющей полосы и минимальной суммарной ошибкой.

По теореме Куна-Таккера сводим задачу к поиску локального минимума заданной функции (седловой точки функции Лагранжа):

$$\begin{cases} -L(\lambda) = -\sum_{i=1}^n \lambda_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j c_i c_j (\mathbf{x}_i \cdot \mathbf{x}_j) \rightarrow \min_{\lambda} \\ 0 \leq \lambda_i \leq C, \quad 1 \leq i \leq n \\ \sum_{i=1}^n \lambda_i c_i = 0 \end{cases}$$

Если выборка может быть разделена на классы линейно, а объекты-выбросы классифицируются неверно, то можно использовать алгоритмы фильтрации выбросов. То есть задача решается при наличии дополнительной переменной - C , при которой из выборки удаляется небольшая часть объектов, имеющих максимальную величину ошибки ξ_i . После таких преобразований задача решается по-новой на усечённой выборке данных. Как правило, требуется проделать некоторое количество итераций до тех пор, пока оставшиеся объекты не смогут быть разделены линейно.

Метод опорных векторов признан одним из самых эффективных классических алгоритмов классификации объектов с использованием моделей классического машинного обучения.

- Алгоритм случайного леса.

Случайный лес — разновидность бэггинг-алгоритмов (частный случай алгоритма усреднения моделей в контексте ансамблевых алгоритмов), который

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	<p>выбросов. То есть задача решается при наличии дополнительной переменной - C, при которой из выборки удаляется небольшая часть объектов, имеющих максимальную величину ошибки ξ_i. После таких преобразований задача решается по-новой на усечённой выборке данных. Как правило, требуется проделать некоторое количество итераций до тех пор, пока оставшиеся объекты не смогут быть разделены линейно.</p> <p>Метод опорных векторов признан одним из самых эффективных классических алгоритмов классификации объектов с использованием моделей классического машинного обучения.</p> <ul style="list-style-type: none">Алгоритм случайного леса. <p>Случайный лес — разновидность бэггинг-алгоритмов (частный случай алгоритма усреднения моделей в контексте ансамблевых алгоритмов), который</p>				
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)			Лист	
								26	

показал свою эффективность в оценке какой-либо статистической величины (например, среднего значения). Входные данные разделяются на множество подвыборок, для которых считается среднее значение, после чего осуществляется усреднение результатов для получения лучшей оценки действительного среднего значения. То есть моделей создается столько же, сколько и подвыборок.

Как правило, для оценки статистических моделей чаще всего используются деревья решений. При построении деревьев решений для создания каждого узла выбираются случайные признаки.

- Бустинг.

Бустинг — это семейство ансамблевых алгоритмов, которые создают один сильный классификатор на основе нескольких слабых. То есть создаётся модель, затем другая модель, базирующаяся на старой, но с некоторыми улучшениями - исправлениями ошибкой в первой (пример: 1.9). Модель обновляется до тех пор, пока тренировочные данные не будут предсказываться настолько точно, насколько это возможно, или пока не будет превышено максимальное число моделей.

Так как весь алгоритм построен на исправлении ошибок моделей, важно, чтобы в данных отсутствовали аномалии, то есть данные должны быть нормализованы и не включать выбросы.

Существует большое количество алгоритмов бустинга, однако только алгоритмы в формулировке приближённо правильного обучения, могут быть точно названы алгоритмами бустинга. Другие алгоритмы, близкие по духу алгоритмам бустинга, иногда называются «алгоритмами максимального использования» (англ. leveraging algorithms), хотя они иногда также неверно называются алгоритмами бустинга.

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	Лист
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)
					27

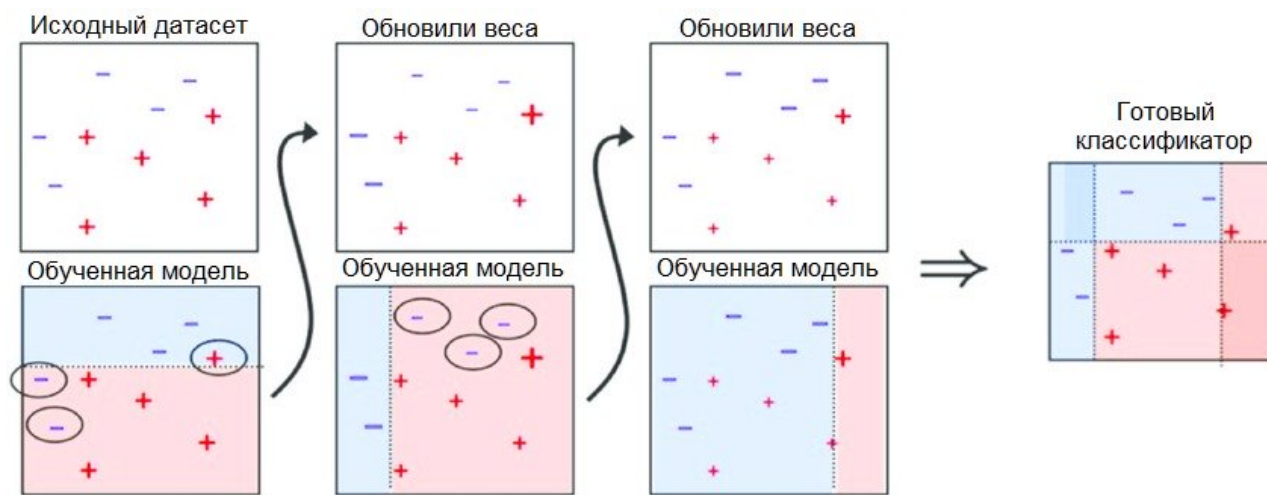


Рисунок 1.9. Иллюстрация работы бустинг-алгоритмов

Основное расхождение между многими алгоритмами бустинга заключается в методах определения весовых коэффициентов точек тренировочных данных и гипотез. Алгоритм AdaBoost очень популярен и исторически наиболее знаменателен, так как он был первым алгоритмом, который смог адаптироваться к слабому обучению. Многие алгоритмы бустинга попадают в модель AnyBoost, это показывает, что бустинг осуществляет градиентный спуск в пространстве функций используя выпуклую функцию потерь.

1.3.3. Обзор метрик, определяющих адекватность модели

Рассмотрев наиболее популярные алгоритмы машинного обучения, рассмотренные в разделе 1.2.2. "Обзор наиболее популярных алгоритмов машинного обучения; а также оценив их достоинства и недостатки, было принято решение о использовании нескольких алгоритмов машинного обучения и сравнении результатов их работы. Существует несколько метрик сравнения результатов работы моделей[10]:

- Ассурасу, Точность классификации - число правильных прогнозов, сделанное как отношение всех сделанных прогнозов.

$$accuracy = \frac{TRUE_{positive} + TRUE_{negative}}{TRUE_{positive} + TRUE_{negative} + FALSE_{positive} + FALSE_{negative}}$$

Данная метрика широко применяется для оценки точности и адекватности модели для задач классификации. Стоит отметить, что данная оценка адекватна тогда и только тогда, когда в каждом классе имеется примерно равное количество случаев, и что все предсказания и ошибки предсказания одинаково важны.

- Классификационный отчет - Precision (точность), recall(полнота) и F-мера
Точность, Precision - это некоторая доля объектов, обозначенных классификатором положительными и при этом действительно являющимися положительными результатами.

$$precision = \frac{TRUE_{positive}}{TRUE_{positive} + FALSE_{positive}}$$

Полнота, Recall - мера, отражающая долю объектов положительного класса из всех объектов положительного класса нашел алгоритм.

$$recall = \frac{TRUE_{positive}}{TRUE_{positive} + FALSE_{negative}}$$

Precision и recall не зависимы от соотношения классов, что означает их применимость в условиях несбалансированных выборок.

F-мера — среднее гармоническое precision и recall :

$$recall = (1 + \beta^2) \frac{recall \cdot precision}{\beta^2 \cdot precision + recall},$$

где β - вес точности метрики.

- Логарифмическая потеря - метрика точности для оценки предсказаний вероятностей принадлежности к классу. То есть это функция, характеризующая потери при неправильном принятии решений на основе наблюдаемых данных.

$$logloss = -\frac{1}{l} \sum_{i=1}^l (y_i \cdot \log(\tilde{y}_i) + (1 - y_i) \cdot \log(1 - \tilde{y}_i)),$$

где l - объем выборки, \tilde{y}_i - ответ алгоритма на i -м объекте, y_i - правильный объект

- Площадь под ROC-кривой

ROC-кривая (receiver operating characteristic) — кривая, оценивающая качество классификации и отображающая соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущие признак, и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак.

AUC (area under ROC curve) — площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор.

- Матрица путаницы (Confusion matrix) - таблица, визуализирующая производительность алгоритма. Каждая строка матрицы представляет экземпляры в прогнозируемом классе, а каждый экземпляр столбца представляет экземпляры в реальном классе (или наоборот).

Наиболее часто используемыми и более информативными метриками являются:

- AUC
- Точность
- Классификационный отчет

Выводы по Главе 1

Проанализировав существующие подходы к решению поставленной задачи, а также их главные преимущества и недостатки, было принято решение о применении статистического метода решения задачи классификации сетевого трафика с

Инва. № подл.	Подп. и дата
Инва. № дубл.	Взам. инв. №
Подп. и дата	
Инва. № подл.	

Ли	Изм.	№ докум.	Подп.	Дат

использованием данных о сетевых сессиях и выделение на их базе поведенческих привычек пользователей.

В данной главе был описан общий подход к решению задач с использованием алгоритмов машинного обучения, а также были описаны наиболее популярные алгоритмы данного класса и меры оценки их адекватности.

Применение данного подхода требует дополнительных инструментов (дополнительного программного обеспечения) и дополнительного анализа и обработки собранных данных, что будет подробно рассмотрено в следующей главе.

Инв. № подл	Подп. и дата				Инв. № дубл.	Взам. инв. №	Подп. и дата	
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)			Лист
								31

2. Информационная модель классификации сетевого трафика

Предлагаемый метод решения поставленной задачи основан на статистических методах. Данный подход занимает относительно малое количество машинного времени, но при этом требует дополнительного анализа и обработки данных. Стоит отметить, что глубокий анализ сетевых пакетов производится на стороне анализатора пакетов, поэтому предлагаемое решение, по сути, является комбинацией существующих подходов для решения задачи классификации сетевого трафика [11]. Признаковое описание будет сформировано на основе характеристик сетевых сессий, из которых будут выведены поведенческие привычки пользователей.

Этапы разработки модели идентификации пользователя по сгенерированному сетевому трафику можно описать схемой 2.1

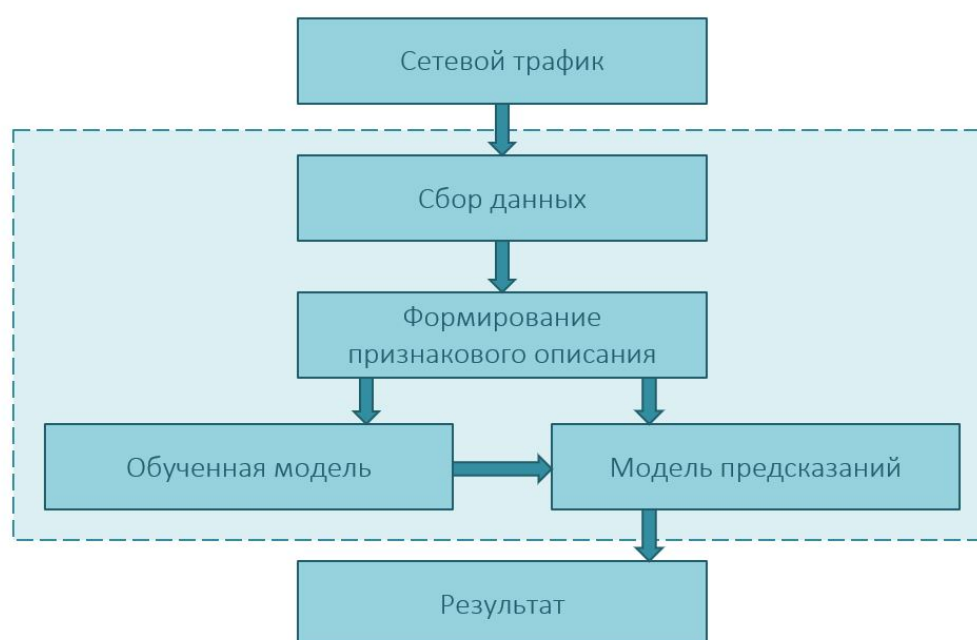


Рисунок 2.1. Шаги подготовки модели классификации сетевого трафика

2.1. Этап сбора данных для алгоритмов машинного обучения

Под сбором данных (в контексте машинного обучения)[9] обычно понимают сбор информации и оценку качества полученных данных, использующихся для обучения выбранной модели, способной решить поставленную задачу и выдать корректный результат, на вновь полученных данных.

Этап сбора данных является одним из самых важных этапов в процессе решения задачи идентификации пользователя по сетевому трафику с использованием моделей и алгоритмов машинного обучения. Именно на этом этапе нужно определить признаковое пространство (правильно разметить данные), а так же выделить и удалить из выборки излишние данные, которые могут только ухудшить качество обучения модели.

Этап подготовки данных для алгоритмов машинного обучения можно разделить на несколько этапов:

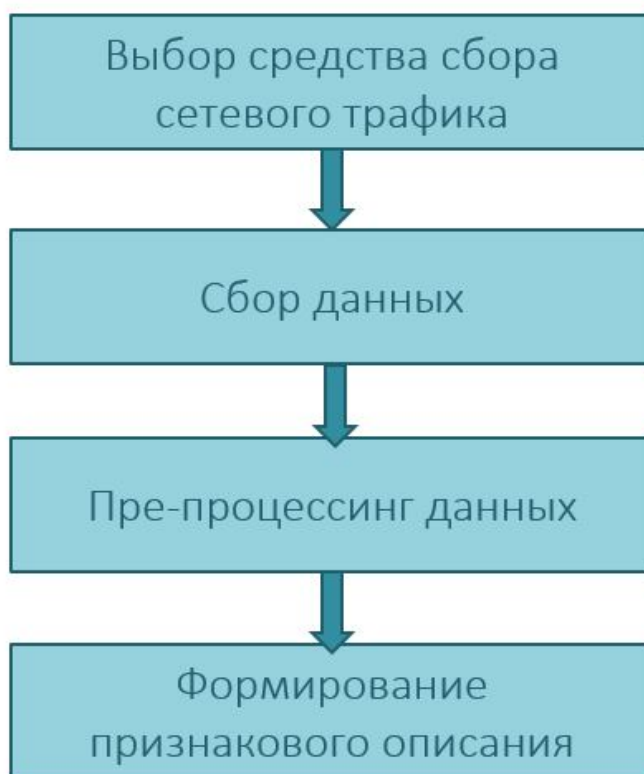


Рисунок 2.2. Этапы подготовки данных для обучения модели

Инв. № подл	Подп. и дата				Лист
	Инв. № дубл.				
	Взам. инв. №				
	Подп. и дата				
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ) 33

```
graph TD; A[Выбор средства сбора сетевого трафика] --> B[Сбор данных]; B --> C[Пре-процессинг данных]; C --> D[Формирование признакового описания];
```

Рисунок 2.2. Этапы подготовки данных для обучения модели

Рассмотрим каждый из приведенных этапов, указанных на схеме 2.6 подробнее.

2.2. Выбор средства сбора сетевого трафика

Существует несколько способов сбора сетевого трафика пользователя[12]:

- Захват сетевых пакетов на устройстве пользователя с использованием специализированного программного обеспечения - анализатора трафика или sniff-фера (программное или аппаратное обеспечение для перехвата и анализа сетевого трафика (своего и/или чужого)).

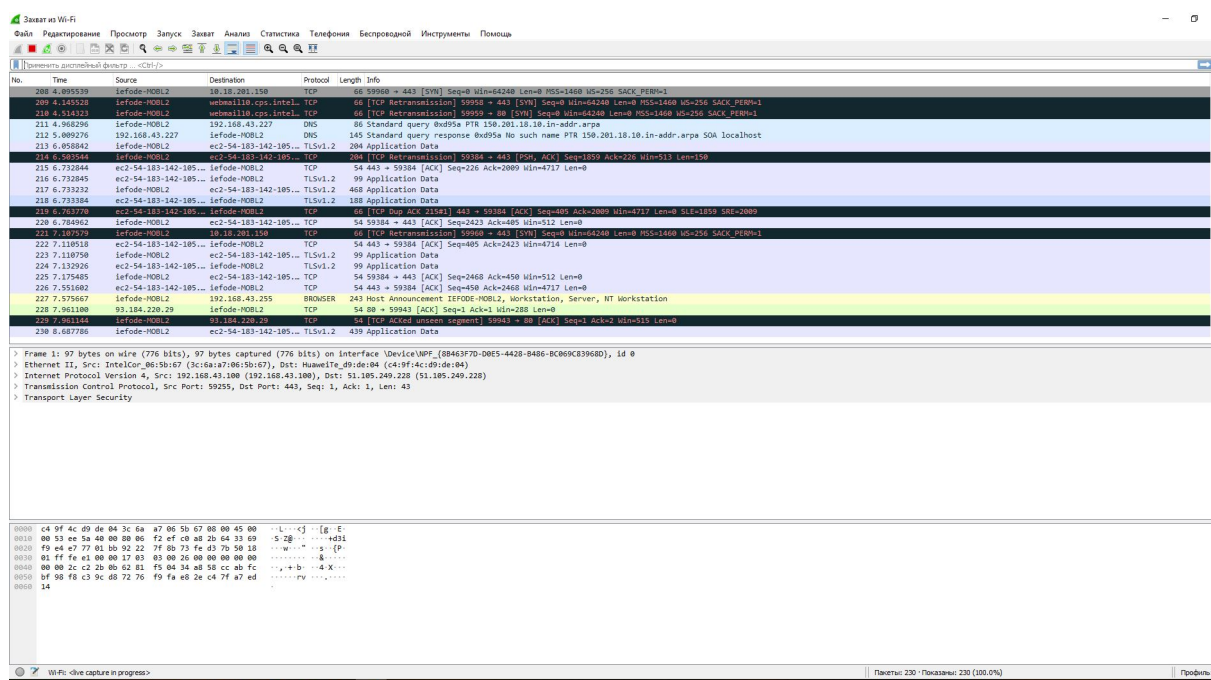


Рисунок 2.3. Скриншот работы sniff-фера Wireshark

Принцип работы заключается в пропускании сетевого трафика через виртуальную или физическую сетевую карту sniff-фера. Сетевые пакеты рассылаются всем машинам в рамках одной сети, благодаря чему становится возможным перехватывать чужую информацию. В качестве защиты от прослушивания можно использовать коммутаторы, грамотно сконфигурировав их - информация между сегментами передается через коммутаторы. Коммутация пакетов — это форма передачи данных, при которой информация

разделяется на отдельные пакеты и передается из исходного пункта в пункт назначения различными маршрутами. Что означает, что рассылка в другом сегменте различных пакетов, не может отправить данные через дополнительный коммутатор. Чтобы начать собирать статистику по использованию

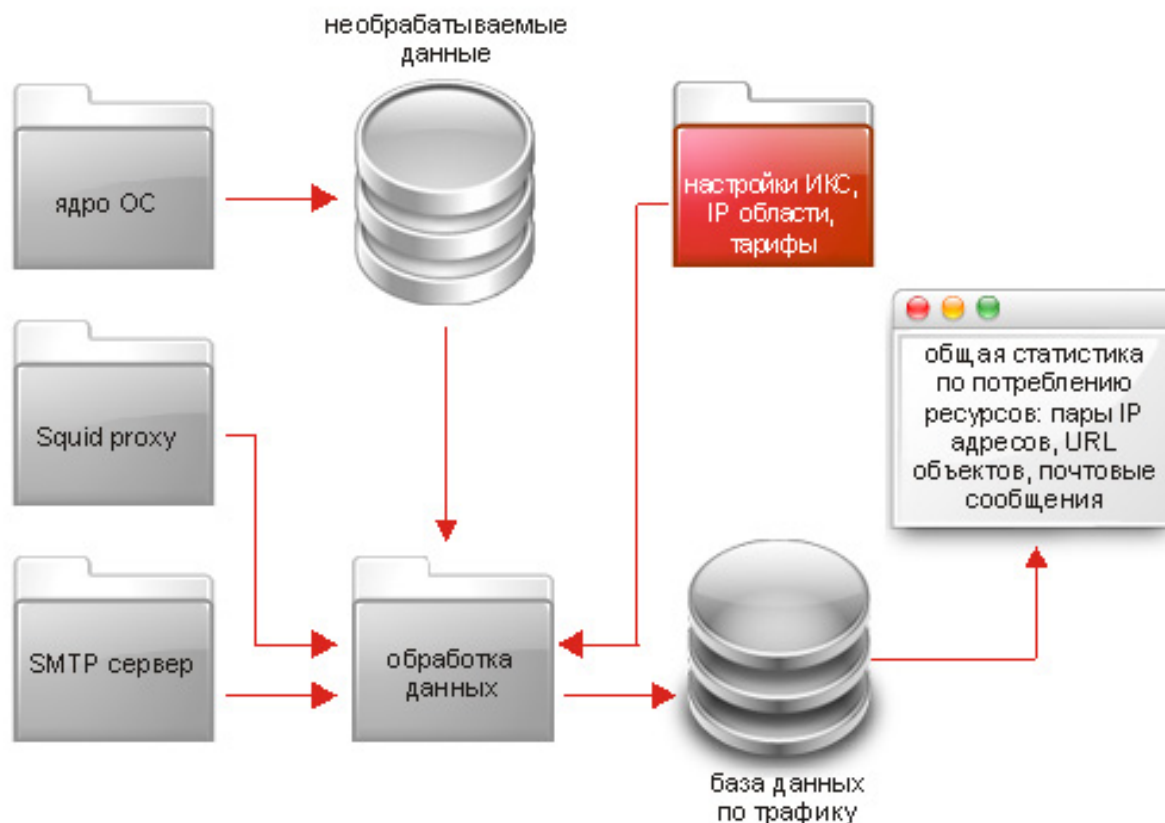


Рисунок 2.4. Принцип сбора трафика через межсетевой экран

сетевого трафика достаточно установить сниффер и запустить его (Например, WireShark). Полученные данные легко можно заэкспортировать в любой поддерживаемый формат (в том числе, табличный или объектного описания). Сами же данные будут содержать в себе всю необходимую для анализа информацию, так как на стороне данного программного обеспечения уже производится глубокий анализ передаваемых пакетов.

Сниффер может быть установлен в нескольких местах:

- прослушивание "сетевого интерфейса"
- подключение сниффера в разрыв сетевого канала

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата
Ли	Изм.	№ докум.	Подп.	Дат
ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)				
				Лист 35

- подключение sniffера в ответвление сети и дублирование сетевого трафика на прослушивающее программное или аппаратное устройство
 - анализ побочных электромагнитных излучений и восстановление из полученных данных сетевого трафика
 - атака на канальном или сетевом уровне. Последствием этого является перенаправление сетевого трафика на sniffer, с последующим возвращением в место назначения.
- Получение статистики использования сетевого трафика с маршрутизатора сети.

<div> <div>Status</div> <div>Quick Setup</div> <div>QSS</div> <div>Network</div> <div>Wireless</div> <div>DHCP</div> <div>Network Sharing</div> <div>Forwarding</div> <div>Security</div> <div>Parental Control</div> <div>Access Control</div> <div>Advanced Routing</div> <div>Bandwidth Control</div> <div>IP & MAC Binding</div> <div>Dynamic DNS</div> <div>System Tools</div> <div>- Time Settings</div> <div>- Diagnostic</div> <div>- Firmware Upgrade</div> <div>- Factory Defaults</div> <div>- Backup & Restore</div> <div>- Reboot</div> <div>- Password</div> <div>- System Log</div> <div>- Statistics</div> </div>	System Log				
	Auto Mail Feature: Disabled Mail Settings				
	Log Type: All Log Level: ALL				
	Index	Time	Type	Level	Log Content
	1	Dec 17 01:34:16	OTHER	INFO	System started
	2	Dec 17 01:34:20	DHCP	NOTICE	DHCP server started
	3	Dec 17 01:34:20	SECURITY	INFO	PPTP Passthrough enabled
	4	Dec 17 01:34:20	SECURITY	INFO	L2TP Passthrough enabled
	5	Dec 17 01:34:20	SECURITY	INFO	IPSEC Passthrough enabled
	6	Dec 17 01:34:20	DHCP	NOTICE	DHCP Send DISCOVER with request ip 0 and unicast flag 0
	7	Dec 17 01:34:20	DHCP	NOTICE	DHCP Recv OFFER from server ac1e0401 with ip ac1e0403
	8	Dec 17 01:34:20	DHCP	NOTICE	DHCP Send REQUEST to server ac1e0401 with request ip ac1e0403
	9	Dec 17 01:34:20	SECURITY	INFO	FTP ALG enabled
	10	Dec 17 01:34:20	SECURITY	INFO	TFTP ALG enabled
	11	Dec 17 01:34:20	SECURITY	INFO	H323 ALG enabled
	12	Dec 17 01:34:20	SECURITY	INFO	RTSP ALG enabled
	13	Dec 17 01:34:21	DHCP	NOTICE	DHCP Recv ACK from server ac1e0401 with ip ac1e0403 lease time 86400
	14	Dec 17 01:34:21	DHCP	NOTICE	DHCP GET ip:ac1e0403 mask:ffff00 gateway:ac1e0401 dns1:ac1e0401 dns2:0 static route:0
	15	Dec 17 01:34:21	DHCP	NOTICE	Dynamic IP(DHCP Client) obtained an IP successfully
	16	Dec 17 01:35:04	DHCP	NOTICE	DHCP Recv REQUEST from D8:5D:4C:8F:46:CB
	17	Dec 17 01:35:05	DHCP	NOTICE	DHCP Send ACK to 192.168.1.100
	18	Dec 17 01:35:08	DHCP	NOTICE	DHCP Recv INFORM from D8:5D:4C:8F:46:CB

Рисунок 2.5. Статистика, полученная с WiFi-роутера

Главным недостатком этого способа сбора информации является зависимость от аппаратного обеспечения, так как далеко не все маршрутизаторы позволяют собирать статистику использования сетевого трафика (потому что не все устройства этого класса поддерживают логирование и сохранение информации). Кроме того, для сбора данных требуются дополнительные знания,

умения и документация, а так же дополнительная обработка результатов полученной статистики разного уровня и приоритета. Так же информации может быть недостаточно, так как сбор трафика может вестись исключительно на сетевом уровне.

Для того, чтобы собрать нужную информацию нужно зайти на адрес веб-интерфейса роутера (или в консоль маршрутизатора) под учетной записью администратора, выставить нужные настройки, собрать данные и сохранить в нужном формате полученные данные.

Стоит также отметить, что недостатком данного способа является сбор статистики всей сети, то есть для того, чтобы получить нужные данные, нужно провести дополнительный отсев нерелевантной информации.

- Межсетевой экран ПК-маршрутизатора

Абсолютное большинство брандмауэров работают исключительно на сетевом уровне, а следовательно собирают статистику обращения на порты, то есть информации для решения поставленной задачи просто может оказаться недостаточно. Дополнительной сложностью является добавочный анализ данных и сохранение в нужном формате.

- Интерфейсы операционной системы или физические сетевые интерфейсы

Как правило, данные устройства собирают лишь статистику передачи данных по определенному каналу, что в контексте данной задачи не представляет никакой ценности.

Исходя из вышесказанного, самым доступным способом сбора информации о статистике использования сетевого трафика пользователем является сбор данных посредством специального программного обеспечения – сниффера. Метод использования - "прослушивание" сетевых интерфейсов.

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	Лист	37
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)	

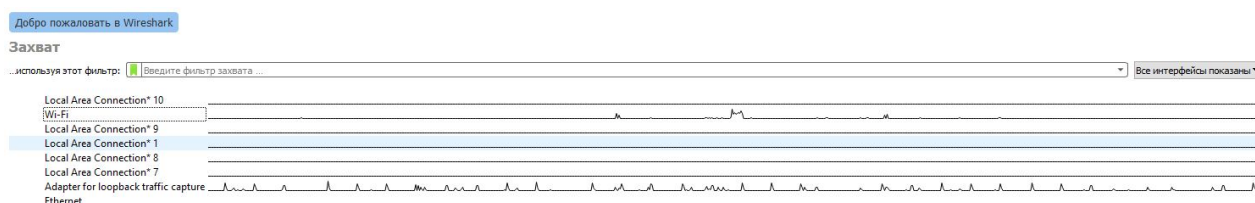
тельно просматривать проходящий по сети трафик в режиме реального времени, переводя сетевую карту в неразборчивый режим (англ. promiscuous mode).

Данное программное обеспечение распространяется под лицензией GNU GPL, позволяющей использовать Wireshark для разработки некоммерческого программного обеспечения, однако, все производное ПО должно быть выпущено под этой же лицензией. Wireshark официально выпускает дистрибутивы для Windows10 и MacOS, однако этот продукт можно собрать на UNIX-подобных системах (GNU/Linux, Solaris, FreeBSD, NetBSD, OpenBSD и другие).

Это приложение, «знающее» структуру разных сетевых протоколов, позволяющее разобрать сетевой пакет и отобразить значение каждого поля протокола всех уровней. Для захвата пакетов используется библиотека pcap, поэтому захватить данные можно только из тех сетей, которые поддерживаются этой библиотекой. Тем не менее, Wireshark поддерживает множество форматов входных данных, поэтому можно открывать файлы данных, захваченных другими программами. Так же в этом продукте обеспечена поддержка скриптового языка Lua.

2.3.2. Сбор данных в приложении Wireshark

Для сбора сетевого трафика с использованием анализатора трафика Wireshark нужно запустить приложение, выбрать локальный сетевой интерфейс для прослушивания (Например, WiFi, Ethernet и другие) и начать захват пакетов.



Локальные сетевые интерфейсы в Wireshark

После сбора сетевого трафика дампы можно сохранить в формате Wireshark, также полученный дампы можно заэкспортировать в любой поддерживаемый фор-

После экспорта содержимого дампа в формате JSON, приведем пример типового объекта пакета, переданному/полученному по стеку TCP/IP:

					ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)
Ли	Изм.	№ докум.	Подп.	Дат	

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата

Ли	Изм.	№ докум.	Подп.	Дат

ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)

- Лист
41

Лист
41

- Лист
41

- Огромные объемы собранной информации даже у небольшого числа пользователей (примерно 5 Гбайт/час при активном использовании трафика)
- Множественное дублирование данных.

2.3.3. Формирование признакового описания данных

Каждый объект обучающей выборки должен иметь признаки, описывающие поведение конкретного пользователя. Проанализировав поля объекта в полученном из анализатора трафика, описывающие переданный/полученный пакет, был сделан вывод, поведение пользователя характеризуется временем и местом соединения, а также характеристиками, описывающими сетевую сессию:

- Имя пользователя.

Будет использоваться как поле, определяющее класс для обучения алгоритмов машинного обучения с учителем

- Страна, из которой было запущено соединение.

Данный параметр можно выделить двумя способами, каждый из которых имеет свои недостатки:

- получить его из описания времени соединения (время соединения и временная зона будет задано настройками используемой операционной системы в поле "Timestamps").

При использовании временной зоны и времени операционной системы, мы не можем утверждать, что пользователь находится именно во временной зоне своей постоянной дислокации, так как операционная система может вообще не ориентироваться на время, заданное сетью.

Пример некорректного значения - пользователь постоянно проживает на территории Нижегородской области, где используется Московский часо-

Инв. № подл	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата	Лист 42
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)

вой пояс (UTC+3 MSK), однако в данный момент временно находится в США.

– вычислить страну по ip-адресу хоста.

При использовании для вычисления местоположения пользователя главным препятствием являются VPN (Virtual Private Networks) и другие похожие решения, которые могут быть организованы в другой стране.

Пример - пользователь является сотрудником Испанской компании, при этом не имеет фиксированного рабочего места и может работать из любой точки мира, и для организации своей постоянной работы использует VPN, организованный в Испании. Таким образом мы будем получать сетевой трафик с IP-адресом испанского сервера, что будет некорректно.

Проанализировав эти случаи, было принято решение использовать временные зоны, заданные операционной системы, так как это даст более полную информацию относительно постоянного местоположения пользователя.

- День недели, в который было открыто соединение.

Это поле может быть получено из поля "timestamp". Его значение это unsigned int число, заданное в UNIX-timestamp, равное количеству секунд, прошедших с полуночи 1 января 1970 года по усреднённому времени Гринвича (т.е. нулевой часовой пояс, точка отсчёта часовых поясов). При получении из базы отображается с учётом часового пояса, заданного в операционной системе, глобальных настройках баз данных или в конкретной сессии. Сохраняется всегда количество секунд по UTC (универсальное координированное время, солнечное время на меридиане Гринвича), а не по локальному часовому поясу.

Инов. № подл	Подп. и дата	Инов. № дубл.	Взам. инв. №	Подп. и дата						
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					Лист
										43

С помощью библиотек временную метку переводим в день недели - число в отрезке от 0 до 6.

- Час и минута, в который было открыто соединение.

Эти данные будут получены и заполнены аналогично предыдущему пункту.

Такого рода данные характеризуют поведение конкретного пользователя, а это значит, что нужно добавить ее в признаковое описание объекта.

- IP-адрес - источник и его имя хоста.

Данная информация важна не только с точки зрения размещения адреса в сети, но и имени машины, с которой было сделано соединение, так как эти данные могут меняться, но при этом характеризовать соединение.

- IP-адрес - назначения и его имя хоста
- TCP-порт-источник
- TCP-порт-назначения

2.4. Выбор алгоритмов машинного обучения для решения поставленной задачи

На основании информации, приведенной в главе 1.2.2. "Обзор наиболее популярных алгоритмов машинного обучения" было принято решение рассмотреть различные алгоритмы для выбора наиболее точной модели, способной идентифицировать пользователя по сгенерированному сетевому трафику:

- Метод опорных векторов
- Наивный байесовский классификатор
- К-ближайших соседей
- Алгоритм случайного леса

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	<ul style="list-style-type: none">• ТСР-порт-назначения
					2.4. Выбор алгоритмов машинного обучения для решения поставленной задачи
					На основании информации, приведенной в главе 1.2.2. "Обзор наиболее популярных алгоритмов машинного обучения" было принято решение рассмотреть различные алгоритмы для выбора наиболее точной модели, способной идентифицировать пользователя по сгенерированному сетевому трафику:
					<ul style="list-style-type: none">• Метод опорных векторов• Наивный байесовский классификатор• К-ближайших соседей• Алгоритм случайного леса
Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	
Ли	Изм.	№ докум.	Подп.	Дат	
ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					Лист 44

- Логистическая регрессия
- Бустинг

Предложенные методы охватывают все наиболее популярные методы решения задач с использованием алгоритмов машинного обучения. Это позволит сравнить результаты работы алгоритмов между собой, что даст более глубокое понимание в выборе наиболее эффективной модели.

2.5. Выбор метрик оценки адекватности моделей

В ходе разработки программного продукта на этапе анализа данных было выявлено, что количество объектов для каждого класса отличается, что означает некорректность работы некоторых метрик. Поэтому было решено произвести оценку адекватности моделей сразу по нескольким признакам - наиболее популярным (см. глава 1.2.3. "Обзор метрик, определяющих адекватность модели").

Перечислим их:

1. AUC;
2. Точность;
3. Классификационный отчет.

Выводы по Главе 2

В данной главе были рассмотрены теоретические основы решения задачи классификации сетевого трафика с использованием алгоритмов машинного обучения - от этапа сбора данных и вплоть до получения результатов. Были выбраны алгоритмы, с помощью которых будет решаться данная задача, а также выбраны меры оценки адекватности работы модели.

Экспериментальные данные о результатах работы предложенных методов на различных наборах исходных данных будут подробно описаны в главе 3.

Инва. № подл.	Подп. и дата
Инва. № дубл.	Взам. инв. №
Подп. и дата	
Инва. № подл.	

Ли	Изм.	№ докум.	Подп.	Дат

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата

Ли	Изм.	№ докум.	Подп.	Дат

ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)

Лист
46

После анализа полученных данных было решено преобразовать данные, каждый объект которых имеет определенную структуру и признаковое пространство, а также принадлежность к классу с последующей целью использования данных для обучения алгоритмов машинного обучения с учителем. Поля были получены как из описания объектов, полученных путем сбора, так и преобразованием полей и их разбиением.

Данные объекты были сведены в csv-таблицу с целью последующего обучения алгоритмов машинного обучения для решения задачи идентификации пользователя по сетевому трафику, которая сводится к решению задачи классификации.

В главе 2.5. Выбор метрик, определяющих адекватность модели был сделан выбор метрик, оценивающих адекватность и точность модели. А в главе 2.6. Выбор алгоритмов машинного обучения для решения поставленной задачи были определены алгоритмы, которые будут использованы для решения поставленной задачи. После проведения всех вычислительных экспериментов, на основании выбранных метрик будет проведена оценка работы каждого алгоритма и сравнение их между собой на основании выбранных метрик.

Главная цель вычислительного эксперимента - найти максимально точный алгоритм для задачи идентификации пользователя на основе сетевого трафика.

3.2. Описание программного продукта

Описанный вычислительный эксперимент являет собой сложный симбиоз сбора данных и их аналитики средствами языков программирования. По сути, вычислительный эксперимент представляет собой цепочку:

1. сбор данных
2. формирование признаковового описания
3. пре-процессинг данных
4. модель идентификации пользователя
5. анализ качества работы модели

Проанализировав различные средства работы с данными, в качестве языка программирования был выбран Python, так как язык предоставляет широкий инструментарий для обработки данных и использования алгоритмов машинного

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата						Лист
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					47

обучения за счет огромного количества пакетов и дополнительных надстроек, которые позволяют собрать целую цепочку из нескольких скриптов и поддерживать так называемый END2END (Exchange-to-exchange) сценарий работы программного продукта, то есть охватить всю последовательность действий с целью решения поставленной задачи. Кроме того, огромным преимуществом данного языка программирования является его кроссплатформенность.

В качестве среды разработки программного кода была выбрана интерактивная оболочка для Python - Jupyter Notebook, предоставляющая широкие возможности интерактивной работы с данными, а также встроенных инструменты отображения результаты работы кода в виде таблиц и графиков. Продукт выпускается под лицензией BSD, что позволяет использовать его без серьезных ограничений.

В качестве программных библиотек для работы с алгоритмами машинного обучения были выбраны:

- Yandex CatBoost (реализует уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга) [15]
- Scikit-learn (библиотека, поддерживающая почти все популярные алгоритмы классического машинного обучения) [16]

Обе библиотеки имеют удобный API и подробную документацию, а также являются библиотеками с открытым исходным кодом (распространяются под лицензиями New BSD и Apache2.0 соответственно, что позволяет использовать их почти без ограничений), кроме того поставляют инструменты, позволяющие проверять адекватность работы моделей и экспорта в некоторые форматы.

Для зачитывания данных была использована библиотека Pandas, предоставляющая широкий набор инструментов подготовки данных для алгоритмов ма-

Инв. № подл	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата	Лист 48
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)

шинного обучения, а также поддерживающая общепринятый формат хранения массивов - NUMPY. Лицензия - BSD.

3.3. Подготовка данных для алгоритмов машинного обучения

Для подготовки данных для алгоритмов машинного обучения нужно произвести с ними некоторые операции:

- Научиться выделять пакеты, переданные через стек протоколов TCP/IP.

Во всех пакетах, переданных посредством стека протоколов TCP/IP, в поле "layers" есть поля "tcp" и "ip". Поэтому все остальные объекты (переданные посредством других стеков протоколов) можно упустить.

- Выделить из множества полей те, которые могут оказать влияние на результат работы классификатора.

Эти данные извлекаются из переданных пакетов посредством Deep Package Inspection алгоритмов и являются основополагающими при сетевом соединении, определяемым стеком протоколов TCP/IP, поэтому именно эти данные и возьмем за основу. Однако наличие только этих данных является недостаточным для обучения алгоритмов с целью идентификации пользователя компьютера.

Так как признаковое описание было сформировано в разделе 2.3.3 Формирование признакового описания данных, то сведем описание каждого объекта к определенным в этой главе полям. После всех преобразований данных и фильтрации, получим объект вида:

```
{
  "hours": 15,
  "minutes": 49,
  "country": "Russia",
  "user": "Olga",
  "weekday": 1,
```

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата

Inspection алгоритмов и являются основополагающими при сетевом соединении, определяемым стеком протоколов TCP/IP, поэтому именно эти данные и возьмем за основу. Однако наличие только этих данных является недостаточным для обучения алгоритмов с целью идентификации пользователя компьютера.

Так как признаковое описание было сформировано в разделе 2.3.3 Формирование признакового описания данных, то сведем описание каждого объекта к определенным в этой главе полям. После всех преобразований данных и фильтрации, получим объект вида:

```
{
  "hours": 15,
  "minutes": 49,
  "country": "Russia",
  "user": "Olga",
  "weekday": 1,
```

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата

Ли	Изм.	№ докум.	Подп.	Дат

ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)

Лист 49

```

"ip.src": "13.83.151.160",
"ip.dst": "192.168.1.195",
"ip.src_host": "fe2cr.update.microsoft.com.akadns.net",
"ip.dst_host": "192.168.1.195",
"tcp.srcport": "443",
"tcp.dstport": "50002"
}

```

После преобразований удобно свести все данные в таблицу, где каждая новая строка - объект, а каждый столбец - тот или иной признак:

name	country	hours	minutes	weekday	ip.src	ip.dst	ip.src_host	ip.dst_host	tcp.srcport	tcp.dstport
Olga	Russia	23	12	3	77.88.21.119	172.20.10.8	mc.yandex.ru	172.20.10.8	443	53443
george	Russia	11	39	0	93.158.162.211	192.168.1.60	s82vla.storage.yandex.net	192.168.1.60	443	39312
iefode	Russia	20	01	6	31.216.145.26	192.168.1.96	gfs270n073.userstorage.mega.co.nz	iefode	443	52639
roman	Russia	14	42	0	178.248.233.33	192.168.1.195	qna.habr.com	68edebd2-8177-4957-93bd-58c9107a9b00.local	443	56548

Такого рода данные уже пригодны для анализа алгоритмами машинного обучения.

Распределение по классам неоднородно, о чем говорит диаграмма 3.1

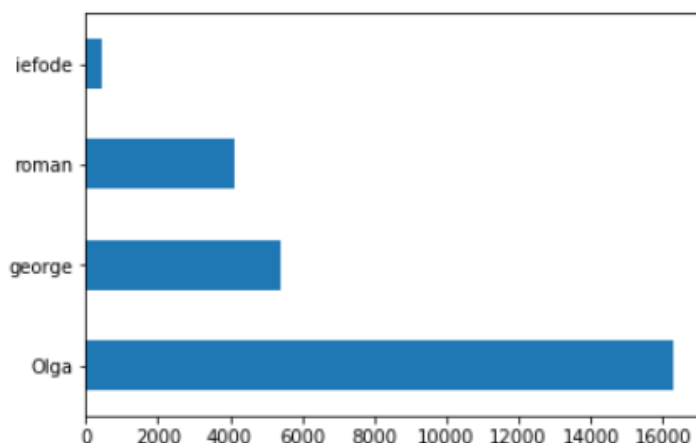


Рисунок 3.1. Диаграмма распределения классов в выборке

Общий объем данных - 28036 объектов (после удаления дубликатов). Из них 70% составляет обучающая выборка, 30% - валидационная. Количество классов - 4 (см. диаграмму 3.1)

3.4. Постановка вычислительного эксперимента

В качестве эксперимента нужно создать различные модели решения задачи классификации с помощью выбранных алгоритмов, а также оценить их адекватность и точность с использованием метрик.

Перед началом работы стоит оценить применимость алгоритмов машинного обучения к данному классу задач.

Матрица корреляции - квадратная таблица, содержащая коэффициенты корреляции между всеми возможными парами переменных, используемых в анализе. Для этого составим матрицу корреляции с использованием библиотеки Pandas на основе собранных данных 3.2:

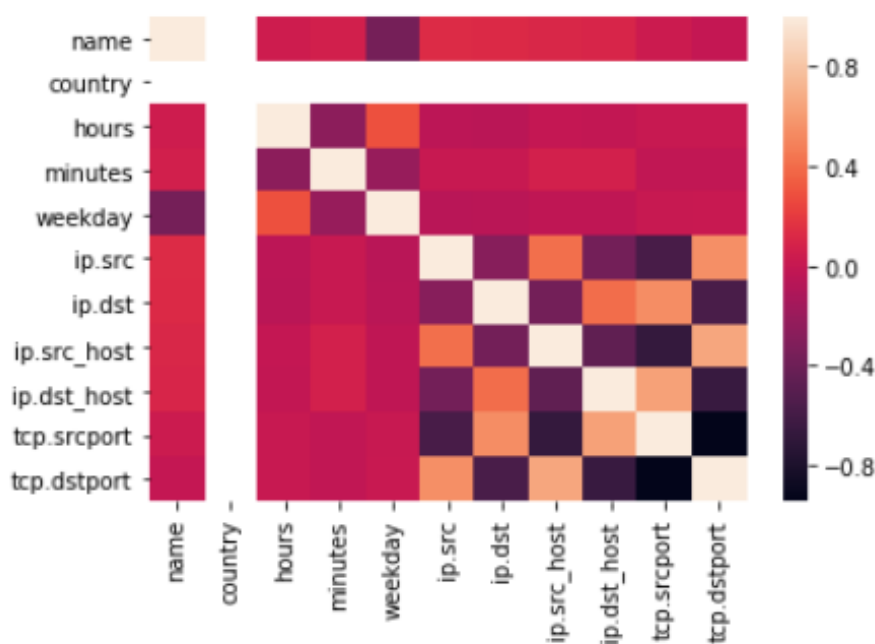


Рисунок 3.2. Матрица корреляции

Как мы можем видеть по рисунку 3.2 многие признаки зависят друг от друга, поэтому алгоритмы машинного обучения могут быть применены для решения данной задачи. Оценим так же насколько каждый из признаков влияет на результат работы алгоритма, то есть насколько важна каждый конкретный признак для предсказания 3.3:

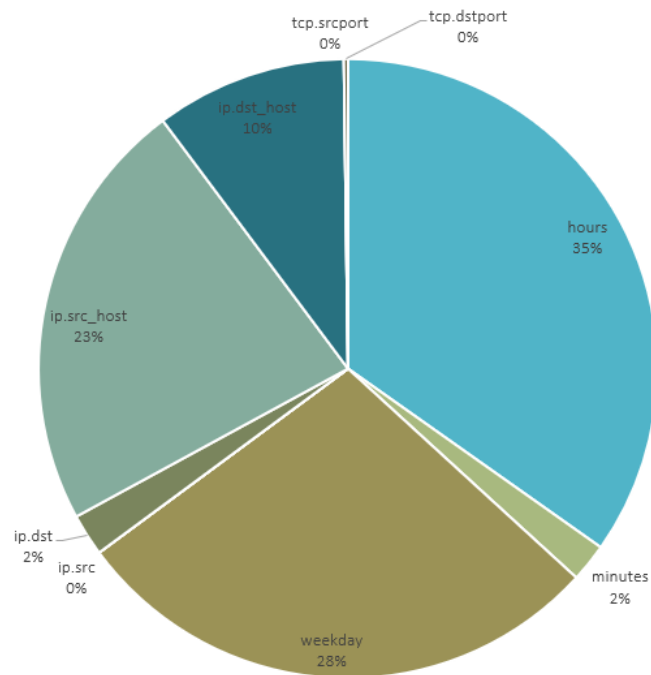


Рисунок 3.3. Диаграмма влияния того или иного признака на результат

Добавим результаты экспериментов:

Accuracy: 0.843419331827				
	precision	recall	f1-score	support
Olga	0.81	0.99	0.89	5190
iefode	0.98	0.63	0.77	1734
roman	1.00	0.29	0.46	139
George	0.94	0.62	0.75	1348
micro avg	0.84	0.84	0.84	8411
macro avg	0.93	0.63	0.71	8411
weighted avg	0.87	0.84	0.83	8411

Рисунок 3.4. Результаты работы SVM нелинейная гиперплоскость - SVC

Accuracy: 0.999881108073				
	precision	recall	f1-score	support
Olga	0.81	0.99	0.89	5190
George	0.98	0.63	0.77	1734
iefode	1.00	0.29	0.46	139
roman	0.94	0.62	0.75	1348
micro avg	0.84	0.84	0.84	8411
macro avg	0.93	0.63	0.71	8411
weighted avg	0.87	0.84	0.83	8411

Рисунок 3.5. Результаты работы SVM - линейная гиперплоскость - LinearSVC

Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата	Лист 52
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)

Accuracy: 1.0					
	precision	recall	f1-score	support	
Olga	1.00	1.00	1.00	5190	
iefode	1.00	1.00	1.00	1734	
roman	1.00	1.00	1.00	139	
George	1.00	1.00	1.00	1348	
micro avg	1.00	1.00	1.00	8411	
macro avg	1.00	1.00	1.00	8411	
weighted avg	1.00	1.00	1.00	8411	

	precision	recall	f1-score	support	pred	AUC
0	1.0	1.0	1.0	5190.0	5190.0	1.0
1	1.0	1.0	1.0	1734.0	1734.0	1.0
2	1.0	1.0	1.0	139.0	139.0	1.0
3	1.0	1.0	1.0	1348.0	1348.0	1.0
avg / total	1.0	1.0	1.0	8411.0	8411.0	1.0

Рисунок 3.6. Результаты работы алгоритмов логистической регрессии

Accuracy: 0.795386993223					
	precision	recall	f1-score	support	
Olga	0.81	0.99	0.89	5190	
iefode	0.98	0.63	0.77	1734	
roman	1.00	0.29	0.46	139	
George	0.94	0.62	0.75	1348	
micro avg	0.84	0.84	0.84	8411	
macro avg	0.93	0.63	0.71	8411	
weighted avg	0.87	0.84	0.83	8411	

	precision	recall	f1-score	support	pred	AUC
0	0.986463	0.758189	0.857392	5190.0	3989.0	0.966263
1	0.955449	0.803922	0.873160	1734.0	1459.0	0.984968
2	0.088254	1.000000	0.162194	139.0	1575.0	0.920394
3	0.880403	0.906528	0.893275	1348.0	1388.0	0.972900
avg / total	0.948227	0.795387	0.854905	8411.0	8411.0	0.934785

Рисунок 3.7. Результаты работы Наивного байесовского классификатора

Accuracy: 0.783497800499					
	precision	recall	f1-score	support	
Olga	0.81	0.99	0.89	5190	
iefode	0.98	0.63	0.77	1734	
roman	1.00	0.29	0.46	139	
George	0.94	0.62	0.75	1348	
micro avg	0.84	0.84	0.84	8411	
macro avg	0.93	0.63	0.71	8411	
weighted avg	0.87	0.84	0.83	8411	

	precision	recall	f1-score	support	pred	AUC
0	0.949514	0.771869	0.851525	5190.0	4219.0	0.852870
1	0.932120	0.768166	0.842238	1734.0	1429.0	0.876819
2	1.000000	0.446043	0.616915	139.0	62.0	0.723022
3	0.440578	0.882789	0.587799	1348.0	2701.0	0.834429
avg / total	0.865197	0.783498	0.803467	8411.0	8411.0	0.855665

Рисунок 3.8. Результаты работы алгоритма k-ближайших соседей

Инв. № подл	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата						Лист
										53
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					

0.6710260373320651					
	precision	recall	f1-score	support	
Olga	0.65	1.00	0.79	5190	
iefode	1.00	0.15	0.25	1734	
roman	0.00	0.00	0.00	139	
George	1.00	0.15	0.26	1348	
micro avg	0.67	0.67	0.67	8411	
macro avg	0.66	0.32	0.33	8411	
weighted avg	0.77	0.67	0.58	8411	
	precision	recall	f1-score	support	pred AUC
0	0.652256	1.000000	0.789534	5190.0	7957.0 1.000000
1	1.000000	0.145905	0.254655	1734.0	253.0 1.000000
3	1.000000	0.149110	0.259522	1348.0	201.0 0.327112
avg / total	0.768899	0.671026	0.581273	8272.0	8272.0 0.955749

Рисунок 3.9. Результаты работы алгоритма случайного леса

Iteration	Learn rate	Test rate	Best score	Total time
0	1.1201049	1.1154156	1.1154156	1.05s
1	0.9391795	0.9307721	0.9307721	2.01s
2	0.8030236	0.7923369	0.7923369	3s
3	0.6928514	0.6816028	0.6816028	3.1s
4	0.6039253	0.5927253	0.5927253	4.11s
...				
145	0.0016911	0.0012560	0.0012560	2m 27s
146	0.0016755	0.0012444	0.0012444	2m 29s
147	0.0016499	0.0012254	0.0012254	2m 30s
148	0.0016280	0.0012100	0.0012100	2m 31s
149	0.0016130	0.0011992	0.0011992	2m 32s
bestTest = 0.001199152998				
bestIteration = 149				

Таблица 3.1. Статистика обучения модели в Yandex Catboost

Отдельно рассмотрим результаты, полученные в ходе работы с библиотекой CatBoost. В ходе обучения была собрана статистика (см. таблицу 3.1).

Как мы видим, самая высокая скорость обучения модели зарегистрирована на начальных этапах, а затем постепенно уменьшается. При этом точность возрастает на каждой итерации. На основании приведенных данных был построен график функции потерь (см. рис. 3.10)

После обучения были получены характеристики модели, отраженные в таблице 3.2.

Loss Function

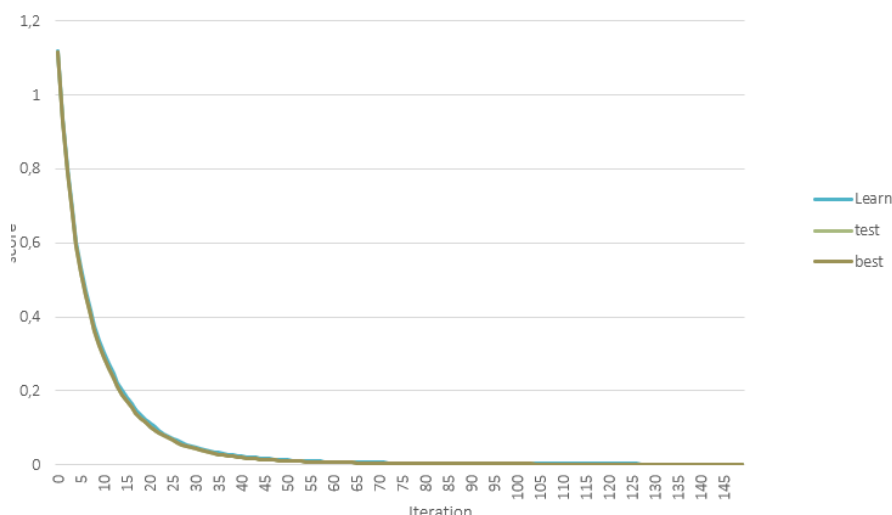


Рисунок 3.10. Функция потерь

	Accuracy	Presicion	Recall	F1	AUC
Olga	1.0	0,9897	0,9914	0,9944	0,9997
george	1.0	0,9979	0,9808	0,9893	
iefode	1.0	1.0	1.0	1.0	
Roman	1.0	0,9946	0,9827	0,9871	
avg	1.0	0,9956	0,9887	0,9923	

Таблица 3.2. Характеристики полученной модели

3.5. Сравнение результатов работы разных алгоритмов

Занесем результаты экспериментов в таблицу 3.3.

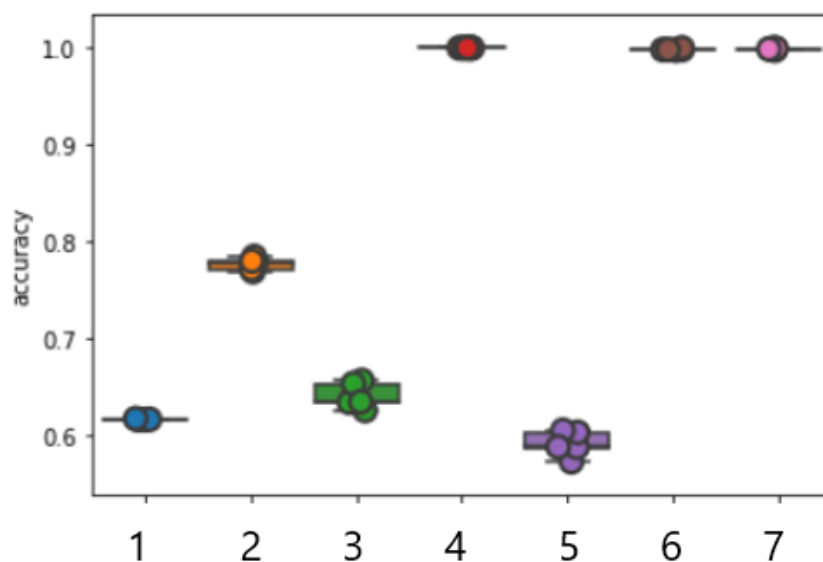
По всем алгоритмам были получены оценки. Для более наглядного и полного сравнения занесем их в единую таблицу 3.3.

Из таблицы можно увидеть, что в контексте задачи идентификации пользователя по сетевому трафику наиболее несостоятельными для решения проблемы

Алгоритм	Accuracy	Precision	Recall	F1	AUC	time
sklearn.svm.SVC	0.8434	0.87	0.84	0.83	0,84	1m28.053s
sklearn.svm.LinearSVC	0.9999	1.0	1.0	1.0	1.0	2.377s
sklearn.linear_model.LogisticRegression	1.0	1.0	1.0	1.0	1.0	0.117
sklearn.naive_bayes.MultinomialNB	0.7954	0.9482	0.7954	0.855	0,9348	2.377s
sklearn.neighbors.KNeighborsClassifier	0,7835	0.8652	0,7835	0,8035	0,8556	0.002s
sklearn.tree.DecisionTreeClassifier	1.0	1.0	1.0	1.0	1.0	0.099s
sklearn.ensemble.RandomForestClassifier	0.6710	0,7689	0,6710	0,5813	0,9557	1.870s
Yandex.CatBoost	1.0	0.9956	0,9887	0,9923	0,9997	2m37s

Таблица 3.3. Результаты работы алгоритмов

можно признать алгоритм (см 3.11) К-ближайших соседей, Наивный байесовский классификатор, и метод опорных векторов с нелинейной гиперплоскостью и алгоритм случайного леса. Максимально точными в контексте применения к решению задачи классификации сетевого трафика можно считать методы логистической регрессии, деревья принятия решений, метод опорных векторов с линейной гиперплоскостью и бустинга. Отдельно стоит отметить алгоритм Yandex.CatBoost, который также зарекомендовал себя для решения данной задачи. Этот алгоритм также содержит внутри себя деревья решений. Из чего можно сделать вывод, что для решения данной задачи подходят алгоритмы, включающие в себя деревья решений.



1. RandomForestClassifier
2. SVC
3. MultinomialNB
4. DecisionTreeClassifier
5. KNNClassifier
6. LinearSVC
7. LogisticRegression

Рисунок 3.11. Диаграмма точности различных алгоритмов машинного обучения для решения задачи классификации сетевого трафика

Инва. № подл	Подп. и дата	Инва. № дубл.	Взам. инв. №	Подп. и дата

Ли	Изм.	№ докум.	Подп.	Дат

Из всего вышесказанного сделаем вывод, что наиболее эффективным показал себя алгоритм с использованием деревьев принятия решений. Он показал отличные результаты по всем метрикам, а также проявил себя как быстро обучающийся.

3.6. Сравнение результатов работы с другими исследованиями

В исследовании австрийских ученых [17, стр. 9], целью которого было детектирование внешних воздействий на систему с использованием различных алгоритмов машинного обучения с целью классификации сетевого трафика были приведены следующие результаты:

<i>Tool</i>	<i>Complexity (Signature)</i>	<i>Training size</i>	<i>Success rate</i>
DirBuster	low	10-50	high (99%)
Burp Suite	none (plain TCP)	5000+	low (40%)
Nessus	complex	5000+	medium (85%)
sqlmap	low	10-50	high (99%)
Nikto	low/medium	10-50	high (95%)

Рисунок 3.12. Результаты работы алгоритмов детектирования внешнего воздействия на систему австрийских исследователей (Метрика Assurance)

В исследовании [18] определяется протокол прикладного уровня с использованием методов глубокого обучения, что сводится к задаче классификации сетевого трафика. В данной работе приведены результаты работы алгоритмов (см. рис. 3.13).

Посчитав среднее получим - $Precision_{AVG} = 0.98$, $Recall_{AVG} = 0.96$.

Исследование [19] посвящено исследованию идентификации сетевого трафика, генерируемого вредоносным программным обеспечением. Проблема, которой посвящена данная работа, так же сводится к решению задачи классификации сетевого трафика с использованием методов, относящихся к обобщенному классу методов искусственного интеллекта. В данном исследовании также дана оценка разным алгоритмам с использованием различных метрик оценки адекватности

Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата	Лист
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)
					57

Результаты, полученные в работах [18, стр. 7], [17, стр. 9], [19, стр. 5] и представленном исследовании сопоставимы между собой по различным метрикам, что говорит о корректности данной работы.

Выводы по Главе 3

В данной главе было описана реализация программного продукта и вычислительного эксперимента, целью которого являлось определение наиболее адекватной модели решения задачи идентификации пользователя по сетевому трафику с использованием алгоритмов машинного обучения и нескольких выбранных метрик и сравнение полученных результатов между собой. Также был подробно описан этап подготовки данных для решения данной задачи.

Был проведен подробный анализ результатов работы предложенных моделей с использованием различных методов оценки адекватности моделей. Кроме этого были выбраны как наиболее точные, так и менее пригодные алгоритмы для решения задачи идентификации пользователя по сетевому трафику.

Исходный код и результаты экспериментов размещены в открытом доступе по адресу:

https://github.com/iefode/traffic_analyzer [20]

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата							
Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата	Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)	Лист 59

Заключение

В ходе работы над выпускной квалификационной работы было выполнено исследование существующих методов и алгоритмов классификации сетевого трафика, в котором были проанализированы различные подходы, а также подсвечены их достоинства и недостатки, изучены проблемы, которые могут возникать при их использовании.

Была проделана работа по сбору данных, в результате которой был определен способ сбора данных и их разметка, а также проведена глубокая аналитика полученных данных, на основании которой был реализован этап подготовки данных для алгоритмов машинного обучения, а также определено признаковое пространство. В ходе анализа существующих подходов к решению задачи идентификации пользователя по сетевому трафику были рассмотрены различные способы решения данной задачи, а также рассмотрены наиболее популярные алгоритмы машинного обучения, приведены их достоинства и недостатки. Кроме того выбраны метрики для оценки адекватности модели.

В ходе реализации программной системы на языке программирования Python, были использованы различные инструменты и подходы. Были реализованы различные вычислительные эксперименты, полученные результаты сравнены между собой, выделены наиболее точные модели. Точность всех алгоритмов была очень высокой - более 75% по всем оцениваемым метрикам.

В дальнейшем планируется создать режиму реального времени с большой базой знаний для идентификации пользователя в режиме онлайн.

Инов. № подл	Подп. и дата	Инов. № дубл.	Взам. инв. №	Подп. и дата
--------------	--------------	---------------	--------------	--------------

Ли	Изм.	№ докум.	Подп.	Дат
----	------	----------	-------	-----

Как результат работы можно признать, что задача идентификации пользователя по сетевому трафику с использованием алгоритмов машинного обучения решена в полном объеме. Кроме того было доказано, что на основании характеристик данных, передаваемых посредством сети, можно идентифицировать конкретного пользователя с достаточно высокой точностью, при этом не изучая содержимое передаваемых пакетов.

Инв. № подл	Подп. и дата				Инв. № дубл.	Взам. инв. №	Подп. и дата	
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)			Лист
								61

Библиографический список

1. Федорук В.Г.: "Протоколы сетевого взаимодействия TCP/IP". МГТУ им. Баумана.
2. Jamuna .A, Vinodh Ewards S.E: "Efficient Flow based Network Traffic Classification using Machine Learning International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, www.ijera.com, Vol. 3, Issue 2, March -April 2013, pp.1324-1328
3. Владимир Хазов: "Классификация трафика и Deep Packet Inspection"[Электронный ресурс]. – <https://vasexperts.ru/blog/klassifikatsiya-trafika-i-deep-packet-inspection/> – Заглавие с экрана. – (Дата обращения: 04.03.20).
4. Bujlow, Tomasz; Riaz, Tahir; Pedersen, Jens Myrup: "A method for classification of network traffic based on C5.0 Machine Learning Algorithm In ICNC'12: 2012 International Conference on Computing, Networking and Communications (ICNC): Workshop on Computing, Networking and Communications (pp. 237-241). IEEE Press. <https://doi.org/10.1109/ICCNC.2012.6167418> Copyright 2011 ACM 978-1-4503-0692-8/11/06
5. Byungchul Park, Young J. Won, Mi-Jung Choi, Myung-Sup Kim², and James W. Hong¹: "Empirical Analysis of Application-level Traffic. Classification using Supervised Machine Learning IT RD program of MKE/IITA [2008-F-016-01,

Инва. № подл	Подл. и дата	Инва. № дубл.	Взам. инв. №	Подл. и дата	ВКР-НГТУ-09.04.01-(М18-ИБТ-3)-006-2020(ПЗ)	Лист
						62
Ли	Изм.	№ докум.	Подп.	Дат		

CASFI] and the EECE division at POSTECH under the BK21 program of MEST, Korea.

6. Fan Zhang, Wenbo He, Xue Liu and Patrick G. Bridges: "Inferring Users' Online Activities Through Traffic Analysis WiSec'11, June14–17, 2011, Hamburg, Germany.
7. Wei Li and Andrew W. Moore: "A Machine Learning Approach for Efficient Traffic Classification EPSRC research grant, GR/T10510/02.
8. Машинное обучение [Электронный ресурс]. – <http://www.machinelearning.ru> – Заглавие с экрана. – (Дата обращения: 04.05.20).
9. Флах П. Машинное обучение. — М.: ДМК Пресс, 2015. — 400 с.
10. I. H. Witten, E. Frank Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). — Morgan Kaufmann, 2005.
11. Mowei Wang, Yong Cui, Xin Wang, Shihan Xiao, and Junchen Jiang: "Machine Learning for Networking: Workflow, Advantages and Opportunites pp 0890-8044/17, IEEE, 2017.
12. Alisha Cecil: "A Summary of Network Traffic Monitoring and Analysis Techniques"[Электронный ресурс]. – https://www.cse.wustl.edu/~jain/cse567-06/ftp/net_monitoring/index.html – Заглавие с экрана. – (Дата обращения: 24.04.2020).
13. Wireshark [Электронный ресурс]. – <https://www.wireshark.org/> – Заглавие с экрана. – (Дата обращения: 10.05.2020).
14. Федеральный закон "О персональных данных" от 27.07.2006 N 152-ФЗ (последняя редакция). Официальный сайт компании "КонсультантПлюс"[Электронный ресурс]. –

http://www.consultant.ru/document/cons_doc_LAW_61801/ – Заглавие с экрана. – (Дата обращения: 10.06.2020).

15. Yandex CatBoost [Электронный ресурс]. – <https://catboost.ai/docs/> – Заглавие с экрана. – (Дата обращения: 21.04.20).

16. scikit-learn. Machine Learning in Python [Электронный ресурс]. – <https://github.com/> – Заглавие с экрана. – (Дата обращения: 29.04.20).

17. P.Frühwirth, S. Schrittwieser, E.R. Weippl: "Using machine learning techniques for traffic classification and preliminary surveying of an attacker's profile".

18. Zhanyi Wang: "The application of deep learning on traffic identification".

19. Alfredo Cuzzocrea, Fabio Martinelli, Francesco Mercaldo, Gianni Vercelli: "Tor Traffic Analysis and Detection via Machine Learning Techniques"

20. GitHub [Электронный ресурс]. – <https://github.com/> – Заглавие с экрана. – (Дата обращения: 19.04.20).

Инв. № подл	Подл. и дата	Инв. № дубл.	Взам. инв. №	Подл. и дата						
Ли	Изм.	№ докум.	Подп.	Дат	ВКР-НГТУ-09.04.01-(М18-ИВТ-3)-006-2020(ПЗ)					Лист
										64