На правах рукописи

Краснова Ирина Артуровна

# Динамическая классификация потоков трафика на основе машинного обучения для обеспечения качества обслуживания в мультисервисной программно-конфигурируемой сети

Специальность 05.12.13 - Системы, сети и устройства телекоммуникаций

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва – 2021

Работа выполнена в ордена Трудового Красного Знамени федеральном государственном бюджетном образовательном учреждении высшего образования «Московский технический университет связи и информатики» (МТУСИ)

**Научный руководитель:** **Деарт Владимир Юрьевич** – кандидат технических наук, доцент, доцент кафедры «Сети связи и системы коммутации» ордена Трудового Красного Знамени федерального государственного бюджетного учреждения высшего образования «Московский технический университет связи и информатики» (МТУСИ)

**Официальные оппоненты:** **Росляков Александр Владимирович** – доктор технических наук, профессор, заведующий кафедрой «Сети и системы связи» федерального государственного бюджетного образовательного учреждения высшего образования «Поволжский государственный университет телекоммуникаций и информатики» (ПГУТИ)

**Гайдамака Юлия Васильевна –** доктор физико-математических наук, профессор, профессор кафедры «Прикладная информатика и теория вероятностей» федерального государственного автономного образовательного учреждения высшего образования «Российский университет дружбы народов» (РУДН)

**Ведущая организация:** Федеральное государственное унитарное предприятие «**Центральный научно-исследовательский институт связи**» (ФГУП ЦНИИС)

Защита состоится «17» февраля 2022 г. в 13:00 часов на заседании диссертационного совета по защите докторских и кандидатских диссертаций Д 219.001.04 (55.2.002.01) при МТУСИ по адресу: 111024, г.Москва, ул. Авиамоторная, д. 8А, МТУСИ, ауд. А-448.

С диссертацией можно ознакомиться в библиотеке и на сайте МТУСИ:

http://www.srd-mtuci.ru/images/Dis-Krasnova/dis-Krasnova.pdf

Автореферат разослан __ ____ 20__ г.

Ученый секретарь

диссертационного совета

Д 219.001.04 (55.2.002.01), д.т.н., доцент Терешонок Максим Валерьевич

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**Актуальность темы исследования.** В последнее время все большее распространение получают программно-конфигурируемые *SDN*-сети (*Software-Defined Networking*), архитектура которых подразумевает отделение плоскости управления от плоскости передачи данных и формирование блока единого централизованного управления сетью. Такая концепция позволяет получить определенные преимущества в организации и управлении информационными потоками, а также добиться большей гибкости в доступе к ресурсам сетей и устройств, и более динамичного развития существующей сетевой инфраструктуры, в отличие от традиционного подхода к построению сетей.

Мультисервисные *SDN*—сети предлагают абоненту целый спектр разнообразных услуг, среди которых выделяют голосовую связь, видеосвязь, видеоконференцию, передачу данных и т.д., а кроме того, позволяют легко добавлять новые приложения и изменять существующие. Каждое из созданных приложений требует обеспечения определенного уровня качества обслуживания *QoS* (*Quality of Service*), что с увеличением числа потребителей и услуг становится все сложнее.

Традиционные механизмы *Differential Services* (услуги с дифференцированным обслуживанием) позволяют обеспечивать определенный уровень *QoS* в рамках одного из классов обслуживания, к которым могут относить тип приложения (*Skype, Telnet* и т.д.), тип передаваемой информации (голос, видео, данные), характер передачи трафика (*Elephant* и *Mice*–потоки - очень большие и очень маленькие потоки соответственно, интерактивный/не интерактивный трафик) и т.д. Для известных оператору сети источников трафика возможна автоматическая или предварительная статическая разметка пакетов с указанием принадлежности сервиса, который используется для дифференцированного управления трафиком. В случаях поступления в сеть пакетов, неучтенных с точки зрения архитектуры сети либо созданных неизвестным для оператора источником (приложением или сервисом), они обрабатываются по принципу негарантированной доставки (*Best Effort*), который не обеспечивает *QoS* на необходимом уровне. Проблема имеет особенное значение для *SDN*-сетей, в которых регулярно появляются новые приложения, а топология сети динамически меняется.

**Степень разработанности темы.**

Предыдущие подходы к классификации потоков в традиционных сетях, основывались на общеизвестном списке *TCP* и *UDP*-портов, но с появлением динамически изменяющихся портов применение такого метода стало невозможным.

Широко известная технология *DPI* (*Deep Packet Inspection*) позволяет проводить «глубокий» анализ заголовков пакетов на верхних уровнях модели ЭМВОС (*OSI*). Но с помощью

системы *DPI* тоже не всегда удается выявить характер потока данных, например, в случаях зашифрованного или туннелированного трафика.

В последнее время в телекоммуникациях все чаще эффективно применяются методы интеллектуального анализа данных, в особенности методы машинного обучения (*Machine Learning, ML*), для решения широкого круга задач, в т.ч. и для классификации трафика.

Исследования и анализ основных работ по классификации методами *ML*, представленные *Гетьманом А.И., Маркиным Ю.В., Ванюшиной А.В., Xie J., Huang N., Zhao D., Perera P., Zhang X., Dong Y., Zhang C., Wang W., Latah M., Pekar A., Habibi L.A., Bakker J.N.* позволили выделить **основные проблемы**, существующие на данный момент:

— проблему с доступом к параметрам пакетов, применяемым в матрице признаков — для большинства работ требуется наличие полной информации о потоке или подробной информации уровня приложений, что означает **ограниченность применения классификатора** для защищенных потоков и **невозможность работы в режиме реального времени**;

— тяжеловесные алгоритмы мониторинга сети, требующие постоянных действий «запрос»- «ответ» от коммутаторов и контроллеров и вызывающие **большую нагрузку** на сеть и сетевые элементы;

— **отсутствие** функциональной **возможности добавления новых классов** в существующую модель классификатора делает невозможным его работу в динамически изменяющихся сетях, таких как SDN.

Российские ученые: *Бурлаков М.Е., Осипов М.Н., Шелухин О.И., Ерохин С.Д.* и зарубежные исследователи: *Alothman B., Gombault S., Toker M., Knapskog S.J., Buczak L., Hodo E., Knapskog S.J., Hamilton A.W., Tachtatzis C., Atkinson R.C.* и др. частично рассматривают эти вопросы, но их **работы сосредоточены** в основном **для** обеспечения **сетевой безопасности**, в то время как классификация трафика с целью определения *QoS* имеет значительные отличия, к которым следует отнести маркировку классов с целью присвоения *QoS*, различия при формировании матрицы признаков и при предварительной обработке данных. Таким образом, разработка методов **классификации** трафика **с целью обеспечения *QoS* требует** проведения исследований и **разработки** других **алгоритмов.**

*Примечание:* здесь и далее под «классификацией трафика» подразумевается «классификация трафика с целью обеспечения *QoS*», если не сказано иное.

**Объектом исследования** являются потоки трафика в мультисервисной *SDN*-сети.

**Предметом исследования** являются характеристики потоков трафика в мультисервисной *SDN*-сети.

**Целью диссертационного исследования** является разработка метода динамической классификации потоков трафика на основе машинного обучения для обеспечения качества обслуживания в мультисервисной *SDN*-сети в режиме реального времени.

**Задачи диссертационного исследования,** решаемые для достижения поставленной цели:

1. Формирование **матрицы признаков** для классификации потоков трафика в режиме реального времени.
2. Разработка **статической модели классификации трафика** методами машинного «обучения с учителем» на основе созданной матрицы признаков.
3. Исследование и разработка **модели кластеризации трафика** методами машинного «обучения без учителя».
4. Создание алгоритма **сбора статистических характеристик** потоков трафика в *SDN*-сетях для сформированной матрицы признаков.
5. Создание эффективного метода **динамической классификации** трафика на основе разработанных моделей, обладающего способностью обнаруживать новые классы и работающего в режиме реального времени.

**Научная новизна.** Научная новизна работы заключается в создании новой модели классификации потоков трафика, отличающейся от ранее представленных следующими ключевыми особенностями:

1. Разработана принципиально **новая матрица признаков**, в которой индивидуальные статистические характеристики каждого из первых 15 пакетов потока предлагается использовать как отдельные признаки, что позволяет применять ее для активных потоков в режиме реального времени, в то время как **общепринятые подходы** предполагают расчет характеристик на основе данных всего потока.
2. Разработана **статическая модель классификации трафика**, **отличающаяся от известных тем, что** включает в себя блок настройки гиперпараметров ансамблевых алгоритмов на основе «решающего дерева» (глубина дерева, количество деревьев, минимальное количество классов в узле для разветвления и т.д.) и блок предварительной обработки данных, основанный на комплексном подходе с использованием квантильной трансформации, параметрического преобразования Йео-Джонсона и работы с выбросами; что позволяет повысить точность классификации и расширить область применения алгоритма *XGBoost* для задач классификации трафика.
3. Доказано **повышение эффективности кластеризации трафика** за счет использования предварительно рассчитанных матриц расстояний на основе методов *Extremely Randomized Trees* и *Random Forest*. В работе **впервые применяется данный подход** для матрицы

признаков режима реального времени, в предыдущих исследованиях кластеризация рассматривается только для применения в моделях вне режима реального времени.

4. Сформирована **система организации памяти** *P4*-коммутатора, организованная за счет выделения ячеек памяти - регистров и назначения им информационного и управленческого функционала. Эта система **отличается от других** своей гибкостью — она позволяет хранить и передавать на контроллер статистическую и идентификационную информацию о любом пакете и только по мере необходимости (например, после накопления первых 10 пакетов), что снижает нагрузку на сеть и устройства по сравнению с традиционными вариантами мониторинга сетевых элементов.

5. Создана **динамическая модель классификации трафика**, полученная за счет объединения точности и скорости работы методов «обучения с учителем» с возможностью образования новых кластеров за счет методов «обучения без учителя», что позволяет ей, **в отличие от других,** обладать одновременно **тремя свойствами**: проводить классификацию для целей обеспечения *QoS*, работать в режиме реального времени и добавлять новые классы в существующую систему.

**Теоретическая и практическая значимость работы.**

Теоретическая значимость исследования обоснована тем, что проведена модернизация существующих математических моделей, позволяющая эффективно применять методы машинного обучения для классификации потоков трафика для обеспечения *QoS* в режиме реального времени; применительно к проблематике диссертации результативно использован комплекс существующих методов машинного обучения; впервые раскрыта и доказана эффективность применения метода *XGBoost* для классификации трафика; изложен метод построения модели классификатора трафика с возможностью добавления новых классов.

Практическую значимость исследования представляют следующие его элементы: статическая модель классификации трафика может применяться на участках сетей с относительно постоянным составом приложений, в т.ч. на традиционных сетях; динамическая модель классификации трафика может применяться для динамически изменяющихся сетей, в т.ч. *SDN*-сетей; метод хранения, сбора и обработки статистической информации может применяться в программируемых *P4*-коммутаторах не только для классификации, но и для других целей мониторинга.

Результаты научно-квалифицированной работы используются при разработке и реализации проектов в **ЗАО «ИнформИнвестГрупп»** и **ООО НПФ «Гранч»**, а также внедрены в учебный процесс кафедры **СиСФС МТУСИ**, что подтверждается соответствующими актами.

**Личный вклад.** Все основные научные положения, промежуточные выводы, представленные в диссертации, получены автором лично.

**Методология и методы исследования.** Для решения поставленных задач применялись методы машинного обучения, математической статистики, теории вероятностей, линейной алгебры, натурного эксперимента и имитационного моделирования.

**Работа соответствует паспорту специальности** 05.12.13 «Системы, сети и устройства телекоммуникаций» по части вопросов комплексного решения технических проблем, задач и вопросов систем и устройств телекоммуникаций. Основные результаты диссертации были получены при работе в следующих областях:

— при исследовании процессов представления информации и создания новых соответствующих алгоритмов и процедур (п. 2) был разработан алгоритм представления данных об информационных потоках трафика в виде матрицы признаков для классификации;

— при разработке эффективных путей развития и совершенствования архитектуры сетей и систем телекоммуникаций и входящих в них устройств (п. 3) была разработана модель классификации трафика с целью поддержания *QoS* в режиме реального времени;

— при разработке новых методов дифференцированного доступа абонентов к ресурсам сетей, систем и устройств телекоммуникаций (п. 5) был предложен новый алгоритм сбора индивидуальной статистической информации о пакетах для *P4*-коммутаторов.

**Положения, выносимые на защиту.**

1. **Матрица признаков** для классификации трафика с целью поддержания *QoS*, в которой признаками являются индивидуальные статистические параметры первых 10-15 пакетов, такие как длина и межинтервальное время прихода пакета на интерфейс, **повышает точность** классификации **на 10-25%** для *TCP* –потоков и до **2%** для *UDP*-потоков по сравнению с другими известными подходами. Структура такой матрицы позволяет эффективно применять классификацию к активным потокам в режиме реального времени и является **инвариантной** по отношению к разным типам потоков трафика.

2. **Статическая модель классификации** трафика на основе созданной матрицы признаков **повышает точность** классификации на **15-25%** для метода **«случайного леса»** и на **30-40%** для **«градиентного бустинга»** по сравнению с другими распространенными подходами за счет использования квантильной трансформации, параметрического преобразования Йео-Джонсона, удаления выбросов и настройки гиперпараметров.

3. **Модель кластеризации трафика**, адаптированная к созданной матрице признаков, **достигает значений согласованного индекса Рэнда 90-100%** за счет применения в процессе кластеризации матрицы расстояний, предварительно рассчитанной на основе результатов классификации потоков методами *Extremely Randomized Trees*. Показано также, что наиболее распространенные и **стандартные подходы** к расчету матриц расстояний, такие как расстояния *Евклида* и *Манхэттена*, **оказались непригодными** к применению в таких условиях.

4. **Алгоритм гибкого сбора статистической информации о пакетах** в *P4*-коммутаторах, основанный на разработанной системе организации памяти, **позволяет хранить и передавать** на контроллер статистическую и идентификационную **информацию о любом пакете** и только **по мере необходимости** (например, после накопления первых 10 пакетов), что **снижает долю** передаваемой **служебной информации**, создающую дополнительную нагрузку на сеть и устройства, в **4-4,5** раза по сравнению с традиционными вариантами полного мониторинга сетевых элементов.

5. Расширенная **динамическая модель классификации** трафика **для целей** обеспечения ***QoS,*** на основе комбинирования методов классификации и кластеризации, **работает в режиме реального времени**, используя скорость и точность работы методов «обучения с учителем», и **добавляет новые классы** за счет методов «обучения без учителя».

**Степень достоверности и апробация результатов.**

Достоверность результатов обеспечивается за счет корректного применения математического аппарата, программного обеспечения и подтверждается результатами расчетов и моделирования.

Основные результаты работы обсуждались на научном межвузовском семинаре *«Современные телекоммуникации и математическая теория телетрафика (СТ и МТТ) № 59»* (Москва, 2021) и **восьми** международных научных конференциях: *«28$^{th}$ Conference of Open Innovations Association (FRUCT)»*, (Москва, 2021), *«The International Science and Technology Conference „Modern Network Technologies, MoNeTec - 2020“»* (Москва, 2020), *«The First International Symposium on Computer Science, Digital Economy and Intelligent Systems (CSDEIS2019)»* (Москва, 2019), *«The Second International Symposium on Computer Science, Digital Economy and Intelligent Systems (CSDEIS2020)»* (Москва, 2020), *«The Fourth International Conference of Artificial Intelligence, Medical Engineering, Education (AIMEE2020)*» (Москва, 2020), *«XI Международной отраслевой научно-технической конференции „Технологии информационного общества“»* (Москва, 2017), *«Международной научно-технической конференции „Информационные технологии и математическое моделирование систем 2019“ (ИТММС 2019)»* (Одинцово, 2019) и «X *Московской научно-практической конференции „Студенческая наука - 2015“*» (Москва, 2015).

По теме диссертации опубликовано **12** печатных работ, из них **5** в научных изданиях, индексированных в международных наукометрических базах, в т.ч. ***WoS***, ***Scopus*** и ***Springer***, **3** в ведущих рецензируемых научных журналах, рекомендованных ***ВАК***, и **1** учебное пособие.

**Объем и структура работы.** Диссертация изложена на 194 страницах, включает в себя 81 рисунок, 33 таблицы и состоит из введения, пяти разделов, заключения, списка источников из 164 наименований, списка сокращений, и 4-х приложений.

## ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

**Во введении** представлено краткое описание актуальности и степени разработанности темы исследования, на основе которых сформулированы цель работы и задачи для ее достижения; выделены основные положения, выносимые на защиту и признаки их научной новизны, за счет которых достигнут искомый положительный эффект; представлен вклад, вносимый в развитие области исследований, а также теоретическая и практическая значимость работы.

**В первом разделе** представлен анализ основных подходов к построению классификации трафика на основе методов машинного обучения. Результаты анализа показали, что большинство работ по классификации посвящено проблемам сетевой безопасности, которые имеют значительные отличия от классификации для целей *QoS*. В работах, посвященных классификации для целей *QoS*, основными проблемами является работа в режиме реального времени и невозможность добавления новых классов в существующую модель, что является серьезным недостатком для *SDN*-сетей. Также отмечены существующие проблемы мониторинга сети для классификации трафика.

Задача классификации трафика может формулироваться следующим образом. Пусть имеется система обслуживания пользовательского трафика (Рисунок 1), на вход которой могут поступать два вида заявок: заявки от размеченных (с помощью меток *DSCP/ToS*, с использованием *VLAN*-тегов и т.д.) потоков: $Y=\{y_1, y_2, \ldots, y_c\}$ и заявки от еще неразмеченных потоков, свойства и природа которых неизвестна: $X=\{x_1, x_2, \ldots, x_k\}$.

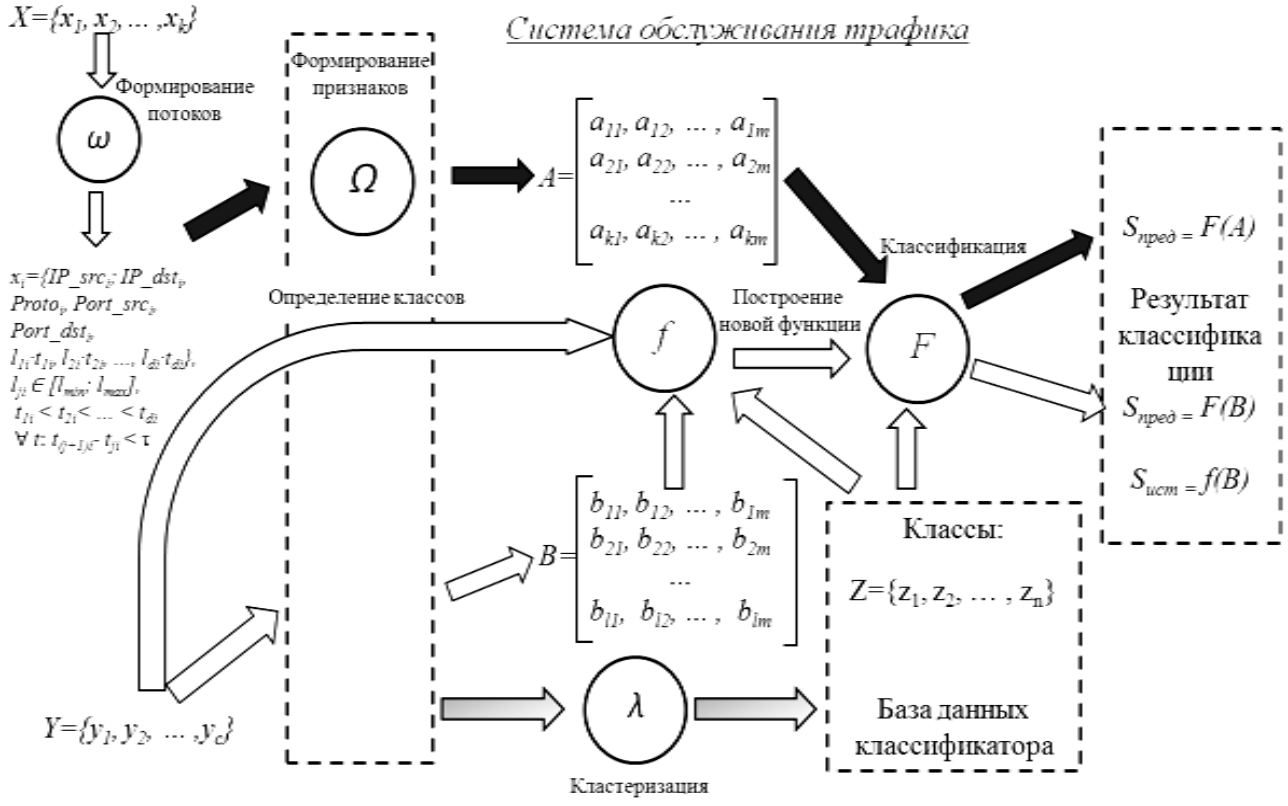



Рисунок 1. - Математическая модель классификации трафика

Поток $x_i$ (или $y_i$) является ординарным и представляет собой последовательные, однонаправлено поступающие пакеты и может быть задан вектором: $x_i=\{IP\_src_i;\ IP\_dst_i,\ Proto_i,\ Port\_src_i,\ Port\_dst_i,\ l_{1i}{\cdot}t_{1i},\ l_{2i}{\cdot}t_{2i},\ \ldots,\ l_{di}{\cdot}t_{di}\}$, где $IP\_src_i$ - IP-адрес источника, $IP\_dst_i$ - IP-адрес назначения, $Proto_{ix}$ - протокол транспортного уровня, $Port\_src_i$ - порт источника и $Port\_dst_i$ - порт назначения - 5-tuple – идентификаторы потока; $l_{ji} \in [l_{min};\ l_{max}]$, $t_{1i} < t_{2i} < \ldots < t_{di};$ – время поступления пакета на интерфейс коммутатора; причем $\forall\ t$: $t_{(j+1)i} - t_{ji} < \tau$ – максимальное межинтервальное время между двумя последовательно поступающими пакетами, определяется из особенностей динамики сети, $l_{min};\ l_{max}$ – минимальная и максимальная длина полезной нагрузки пакета, зависящие от настроек сети и *MTU* (*Maximum Transmission Unit*, максимальная единица передачи). В диссертации исследования проводятся на потоках, в которых $\tau \in (0;\ 180]$ с, $l \in [40;\ 100]$ байт.

Каждый из потоков пространства $X$ и $Y$ с помощью функции выделения признаков $\Omega$ может быть описан своим вектором признаков: $\Omega(X)=A=\{A_1,A_2,\ldots A_m\}$, $\Omega(Y)=B=\{B_1,B_2,\ldots B_m\}$, так что $A_1=\{a_{11},\ a_{21},\ \ldots\ a_{k1}\}$, $B_1=\{b_{11},\ b_{21},\ \ldots\ b_{c1}\}$ и т.д., а $x_i=\{a_{i1},a_{i2},\ldots a_{im}\}$, где $i=1,2,..k$; $y_j=\{b_{j1},\ b_{j2},\ b_{j3},\ldots b_{jm}\}$, где $j=1,2,\ldots,c$. Здесь $m$ – общее число признаков для пространств $X$ и $Y$. В качестве признаков используются статистические характеристики потоков трафика.

Под классом трафика подразумевается определенный тип трафика, обусловленный его особенностями (тип приложения, тип сервиса, укрупненные категории и т.д.). В модели существует некоторая известная база данных классов трафика - пространство $Z=\{z_1,\ z_2,\ \ldots, z_n\}$, где $n$– число известных классов. При этом для пространства $Y$ существует функция $S_{ист}=f(B)=Z$ - «истинные классы», однозначно определяющая соответствие между векторами признаков потоков трафика из пространства $Y$ и классами из пространства $Z$. Для режима **обучения** в качестве $S_{ист}$ используют данные о разметке потоков трафика. Также предполагается существование некой функции $F$, такой, что $S_{пред}=F(B)\approx Z$ – предсказанные классы, при этом функция $F$ изначально неизвестна и не может быть получена с использованием разметки трафика для пространства $Y$.

На этапе разработки модели, до введения системы обслуживания в эксплуатацию, для оценки работоспособности алгоритма часть потоков трафика из пространства $Y$ не отправляется на этап определения функции $F$, вместо этого над ними применяется сама функция $F$: $S_{пред}=F(B)$ и результат сравнивается с $S_{ист}=f(B)$. В терминах машинного обучения этот этап называется **тестированием**.

В результате тестирования определяются основные оценки работы классификатора (1): ***общая точность*** $ACCURACY$ – доля верно классифицированных потоков из всех; ***точность*** каждого ***класса*** $PRECISION$ – доля верно определенных потоков класса из всех потоков, которые были отнесены к этому классу; ***полнота*** каждого класса $RECALL$ – доля верно предсказанных

потоков из всех, принадлежащих этому классу, и ***F1-мера*** – гармоническое среднее между ***полнотой*** и ***точностью*** каждого ***класса***. Здесь $TP_i$ - число верно классифицированных потоков, $F_{ij}$ – число неверно классифицированных потоков, причем $i$ – номер предсказанного класса, а $j$ – номер истинного класса, а $ACCURACY, PRECISION, RECALL, F1 \in [0; 1]$.

$$ACCURACY = \frac{\sum_{i=1}^{n} TP_{ii}}{\sum_{i=1}^{n} TP_{ii} + \sum_{i=1}^{n}\sum_{j=1}^{n} F_{ij}}; \qquad PRECISION_i = \frac{TP_{ii}}{TP_{ii} + \sum_{j}^{n} F_{ij}};$$

$$RECALL_j = \frac{TP_{ii}}{TP_{ii} + \sum_{i}^{n} F_{ij}}; \qquad F1 = \frac{2 \cdot PRECISION \cdot RECALL}{PRECISION + RECALL}. \quad (1)$$

Для добавления новых классов в существующую модель, потоки *X* и *Y* с помощью функции ***λ(X, Y)***$\approx Z$ разбиваются на классы $U = \{U_1, U_2, \dots, U_R\}$ (истинные значения) и кластеры $V = \{V_1, V_2, \dots, V_C\}$ (предсказанные значения). Тогда для числа пар потоков $N_{uv}$, оказавшихся при $u,v=0$ в одном множестве $U$ или $V$ соответственно, а при $u,v=1$ – в разных, согласие истинной и предсказанной выборки определяется через ***согласованный индекс Рэнда*** (*ARI, Adjusted Rand Index*), $ARI \in [-1; 1]$ (2):

$$ARI\ (U,V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}. \quad (2)$$

Пусть $H(U)$, $H(V)$, $H(V|U)$ – энтропия для множеств *U*, *V* и условная энтропия соответственно. Тогда $Hom, Comp, V_{measure} \in [0; 1]$ – ***однородность***, ***полнота кластеров*** и ***V-мера*** (гармоническое среднее однородности и полноты кластеров) являются дополнительными оценками кластеризации (3):

$$Hom = 1 - \frac{H(U|V)}{H(U)}; \qquad Comp = 1 - \frac{H(V|U)}{H(V)}; \qquad V_{measure} = 2 \cdot \frac{Hom \cdot Comp}{Hom + Comp}. \quad (3)$$

Для построения такой модели классификации требуется решить следующие задачи:

1. Разработка матрицы признаков для расчета вектора *A*, т.е. нахождение функции ***Ω(X)***, при этом максимизируются $ACCURACY,\ F1 \to max$.

2. Разработка функции ***F*** с использованием пространства *Y*, *Z* и функции *f*. Функция *F* должна отображать пространство признаков неразмеченных заявок *X* в пространство классов *Z*: $S_{пред}=F(A)\approx Z$, при $ACCURACY,\ F1 \to max$. В терминах машинного обучения – это **задача классификации методами «обучения с учителем»**.

3. Определение базы данных классификатора *Z* на основе пространств *X* и *Y*: ***λ(X, Y)*** $\approx Z$, (при $ARI \to max, V_{measure} \to max$) или **задача кластеризации методами «обучения без учителя»**.

4. Разработка алгоритма сбора статистических характеристик потоков трафика в *SDN*-сетях для формирования пространств *X* и *Y*: ***ω(X, Y)***, минимизирующего долю передаваемой служебной информации $\alpha \to min$.

5. Создание модели классификации трафика с возможностью добавления новых классов на основе композиции разработанных функций: $\boldsymbol{G} = \boldsymbol{\omega} \circ \Omega \circ \boldsymbol{\lambda} \circ \boldsymbol{F}$.

Результаты первого раздела докладывались на конференции *«Технологии информационного общества»* (2017) и опубликованы в статьях [6-7; 10].

**Во втором разделе** разрабатывается матрица признаков для классификации трафика в режиме реального времени с целью поддержания *QoS*. Стандартные подходы к созданию матрицы признаков основаны на подсчете статистических характеристик, таких как средний размер пакета, среднее время поступления между пакетами и т.д. среди всего потока. Для классификации активных потоков в режиме реального времени требуется матрица признаков, доступная по небольшому числу первоначальных пакетов.

Предлагается матрица признаков, в которой признаками являются индивидуальные пакетные характеристики: размеры и межинтервальное время каждого из первых 15-и пакетов (набор признаков 1), а также общепринятые признаки, такие как СКО, дисперсия, суммарное, среднее, минимальное, максимальное, медианное значение размера и межинтервального времени пакетов, пакетная и байтовая скорости (набор признаков 2). Небольшое число пакетов для классификации в диссертации объясняется требованием задачи – результат классификации должен быть получен достаточно быстро, чтобы успеть применить методы управления трафиком к классифицированным потокам.

На данном этапе применялся метод *Random Forest, RF* (*«Случайный лес»*) как наиболее перспективный алгоритм в научных работах по классификации трафика. Метод *Decision Tree, DT* (*«Решающего дерева»*) – логическая модель, состоящая из узлов, листьев и ветвей. В каждом узле находится некоторая условная функция, построенная на основе одного из признаков и разделяющая данные на разные направления, которые, в конечном счете, приходят к одному из листьев, ассоциированных с определенными классами. Метод *Random Forest, RF* представляет собой некоторое количество таких *DT*, которые строятся и принимают решение независимо друг от друга, а итоговый результат классификации определяют по итогам голосования большинства деревьев. Для каждого узла функция разбиения $Q$ множества потоков $S$, размером $N$, на множества левого $S_{left}$ и правого $S_{right}$ узлов, размерами $N_{left}$ и $N_{right}$ соответственно, подбирается таким образом, чтобы минимизировать примеси *G*, оцениваемые по формуле (4):

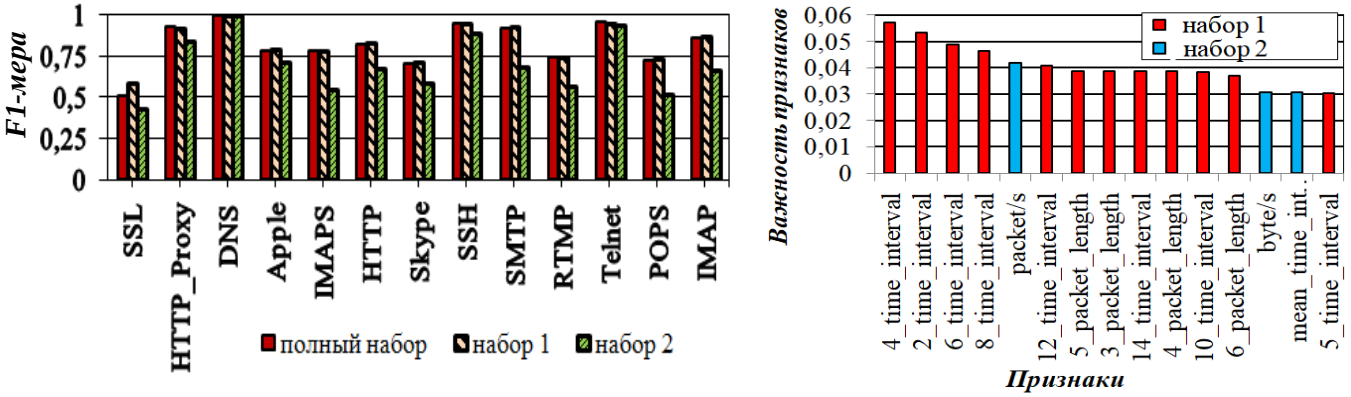
$$G(Q,S) = \frac{N_{left}}{N} \cdot H\left(S_{left}\right) + \frac{N_{right}}{N} \cdot H\left(S_{right}\right) \rightarrow min. \quad (4)$$

Здесь *H* – критерий функции примеси, который рассчитывался на основе индекса Джинни для каждого из узлов *side*, *t* - число посторонних классов, попавших в новый узел, $Z_i^{\text{пост}}$ - *i*-й посторонний класс - $Z_j^{\text{набл}}$ – наблюдаемый *j*-й класс:

$$H(S_{side}) = \frac{1}{N_{side}^2} \cdot \sum_{i=0}^{t} \left( \sum_{j=0}^{N_{side}} I(Z_j^{\text{набл}} = Z_i^{\text{пост}}) \cdot \left( N_{side} - \sum_{j=0}^{N_{side}} I(Z_j^{\text{набл}} = Z_i^{\text{пост}}) \right) \right). \quad (5)$$

Оценка эффективности работы матрицы признаков проводилась на открытой базе данных трафика, собранной в 2020 г. на реальной сети исследовательской группой *MAWI* в рамках проекта *WIDE* в г.Токио. В результате обработки выбранных трасс трафика было выделено два набора данных среди наиболее распространенных приложений: *TCP* (по 1500 потоков из 13 приложений: *SSL, HTTP_Proxy, DNS, Apple, IMAPS, HTTP, Skype, SSH, SMTP, RTMP, Telnet, POPS, IMAP*) и *UDP* (по 250 потоков из 8 приложений: *DNS, NTP, Quic, IPsec, SNMP, NetBIOS, STUN, UPnP*) потоки.

Результаты классификации *TCP* трафика методом *Random Forest* для набора 1, набора 2 и для полного набора из признаков 1 и 2 представлены на Рисунке 2. Оценка ***F1-меры*** классификации (Рисунок 2 (а)) полным набором признаков и набором признаков 1 превышает оценку ***F1-меры*** классификации общепринятым набором признаков 2 для всех приложений, в некоторых случаях прирост ***точности*** классификации доходит до 20% для *TCP* и 2% для *UDP*. Оценка важности признаков (Рисунок 2 (б)) также подтверждает важность признаков предложенного набора.




а) *F1*-мера б) Наиболее важные признаки

Рисунок 2. – Результаты классификации TCP-приложений при разных матрицах признаков

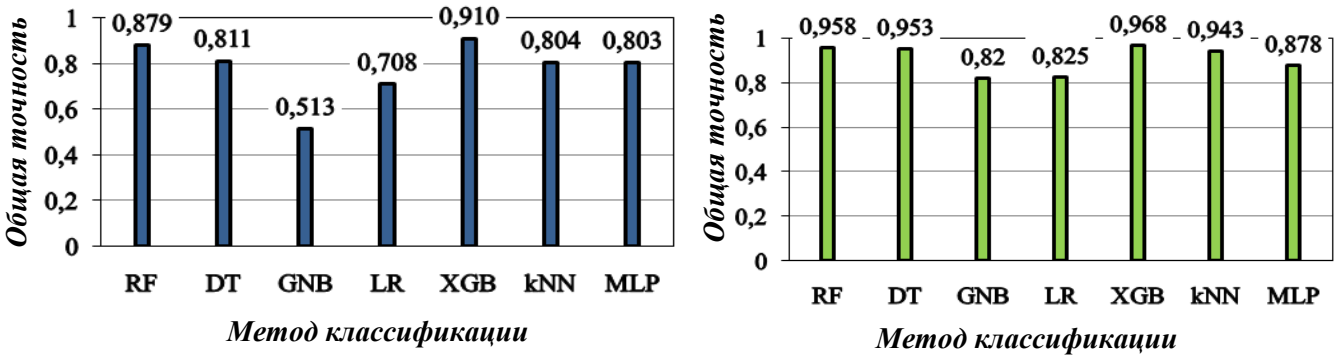
Основные результаты второго раздела обсуждались на конференции *«MoNeTec–2020»* и подробно описаны в статье [2].

**В третьем разделе** создается алгоритм классификации трафика в режиме реального времени для сетей с постоянным составом приложений (Рисунок 6, блоки 2-4). Для сравнительного анализа алгоритмов классификации трафика рассматривались методы: *«Решающее дерево»* (*Decision Tree, DT*), *«Случайный лес»* (*Random Forest, RF*), *«Экстремальный градиентный бустинг»* (*XGBoost, XGB*), *«Гауссовский Наивный Байес»* (*Gaussian Naive Bayes, GNB*), *«Логистическая регрессия»* (*Logistic Regression, LR*), *«k ближайших соседей»* (*k Nearest Neighbours, kNN*) и *«Нейронные сети: Многослойный перцептрон»* (*Neural network: Multi-layer Perceptron, MLP*). При классификации по созданной модели наилучшие результаты показали

*XGBoost* и *Random Forest* (Рисунок 3), причем для построения высокоточной модели *TCP*-трафика достаточно около 1200-1300 потоков для обучения (*XGB*-0,90, *RF*-0,87), а *UDP* – около 200 (*XGB*-0,97, *RF*-0,95). Дальнейшие выводы в автореферате представлены только для *RF* и *XGB*, как наиболее перспективных методов классификации.

*XGB*, как и *RF* является ансамблевым алгоритмом, и в работе также в качестве базовых алгоритмов использовал *DT*, но в отличии от *RF*, *DT* строятся не параллельно, а последовательно, на каждом шаге $\theta$ исправляя ошибки друг друга с помощью целевой функции оптимизации бустинга $obj$ (6), где $\sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t)}\right)$ - специфическая функция потерь, оцениваемая как СКО между значением *i*-го элемента обучающей выборки $y_i$ и предсказания для первых *t* деревьев $\hat{y}_i^{(t)}$; $\frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ – функция контроля суммы весов листьев модели $w_j^2$ с общим числом листьев *T*, а $\gamma$ и $\lambda$ – параметры регуляризации:

$$obj(\theta) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t)}\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2. \quad (6)$$




а) *TCP* – трафик б) *UDP*-трафик

Рисунок 3. - Наилучшие показатели ***точности*** при разных методах машинного обучения

В блоке предварительной обработки данных анализировались результаты классификации при проведении процедур (7-9), применяемых при решении схожих задач или рекомендованные *ГОСТ Р ИСО 16269-4-2017 «Статистическое представление данных. Часть 4. Выявление и обработка выбросов»*, а именно: замена выбросов на медианное значение, при этом в качестве выбросов считаются выборки, превышающие значение СКО (***out***, $a_{ij}^{out}$) и тройное СКО (***out, 3STD***, $a_{ij}^{3out}$), масштабирование данных в пределах [0;1] (***minmax***, $a_{ij}^{MinMax}$) и в пределах между 25 и 75-м квантилем (***robust***,$a_{ij}^{Robust}$); стандартизация по всем данным (***standard***, $a_{ij}^{standart}$); нормализация по каждой выборке отдельно (***normalizer,*** $a_{ij}^{normal}$); трансформация с помощью степенных параметрических преобразований Йео-Джонсона (***power***, $a_{ij}^{power}$), отображающие данные на нормальное распределение и трансформация с помощью отображения кумулятивной функции распределения на равномерное распределение Гаусса (квантильная трансформация, ***quantile,*** $G^{-1}$).

$$a_{ij}^{out} = \begin{cases} a_{ij}, & \forall \ a_{ij} \le a_{ij} \pm \sigma_i \\ M_i, & \forall \ a_{ij} \ge a_{ij} \pm \sigma_i \end{cases}; \qquad a_{ij}^{3out} = \begin{cases} a_{ij}, & \forall \ a_{ij} \le a_{ij} \pm 3\sigma_i \\ M_i, & \forall \ a_{ij} \ge a_{ij} \pm 3\sigma_i \end{cases}; \qquad a_{ij}^{normal} = \frac{a_{ij} - M_i}{\sigma_i}; \quad (7)$$

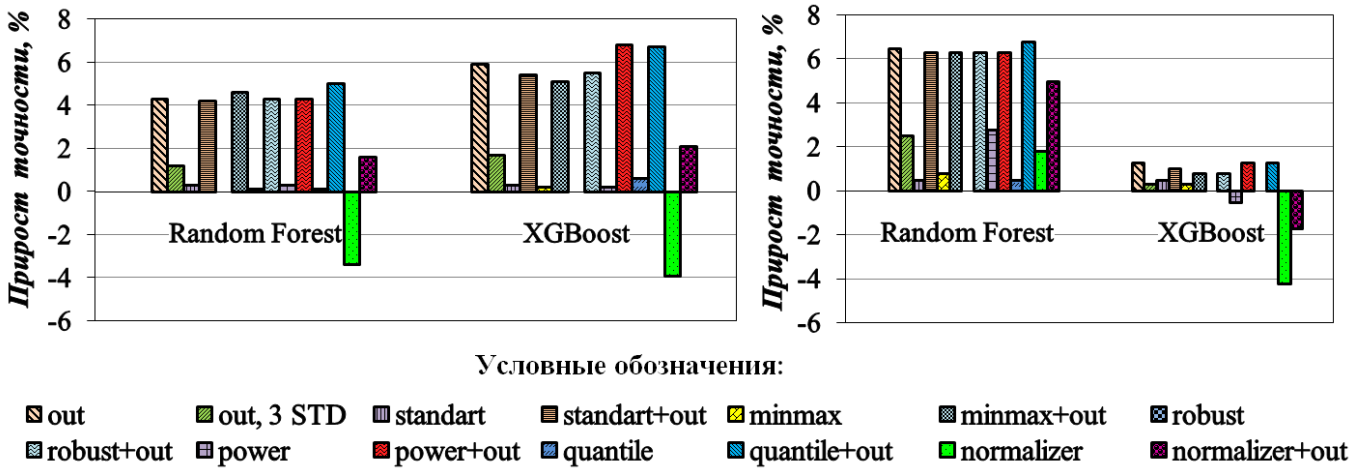
$$a_{ij}^{standart} = \frac{a_{ij} - M}{\sigma}; \qquad a_{ij}^{MinMax} = \frac{a_{ij} - a_{min}}{a_{max} - a_{min}}; \qquad a_{ij}^{Robust} = \frac{a_{ij} - Q_1(A)}{Q_3(A) - Q_1(A)}; \quad G^{-1}(F(a_{ij})); \quad (8)$$

$$a_{ij}^{power} = \begin{cases} \dfrac{\left[(a_{ij}+1)^{\lambda} - 1\right]}{\lambda}, & \text{при } \lambda \ne 0, \ \ a_{ij} \ge 0, \\ \ln(a_{ij}+1), & \text{при } \lambda = 0, \ \ a_{ij} \ge 0, \\ -\dfrac{\left[(-a_{ij}+1)^{2-\lambda} - 1\right]}{2-\lambda}, & \text{при } \lambda \ne 2, \ \ a_{ij} < 0, \\ -\ln(-a_{ij}+1), & \text{при } \lambda = 2, \ \ a_{ij} < 0. \end{cases}; \quad (9)$$

$$l_n(\theta|a_{ij}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(a_{ij}^{power} - \mu\right)^2 + (\lambda - 1)\sum_{i=1}^{k}\sum_{j=1}^{m} sgn(a_{ij}) \cdot \ln(|a_{ij}| + 1).$$

Для формул (7-9): $a_{ij}$ - элемент матрицы признаков $A$, $\sigma_i$, $M_i$, $\sigma$, $M$ – построчное СКО, медиана, полное СКО и медиана; $Q_1(A)$, $Q_3(A)$ – 1-й, 3-й квартили, $\lambda$ – параметр распределения, определяемый из функции максимального правдоподобия $l_n(\theta|a_{ij})$, $n = k \cdot m$ - число элементов матрицы $A$ (Рисунок 1).

Наибольшего прироста ***точности***, которого удалось добиться: для *RF* – 5% при комбинации методов ***out*** и квантильной трансформации ***quantile***; для *XGB* – 7% при комбинации методов ***out*** и степенных параметрических преобразований Йео-Джонсона ***power*** (Рисунок 4).

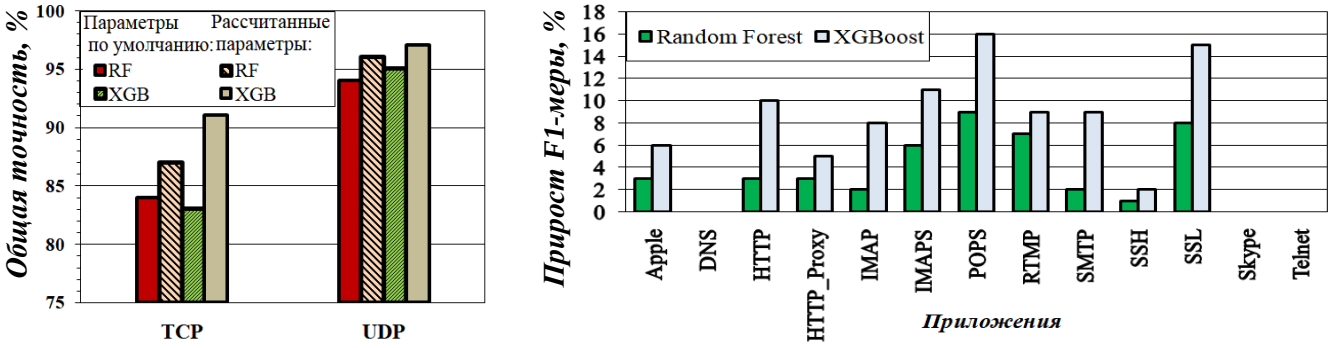



а) *TCP*–трафик б) *UDP*-трафик

Рисунок 4. – Прирост ***точности*** классификации при разных методах предварительной обработки

В блоке классификации проводятся настройки гиперпараметров модели, основная задача которых – борьба с переобучением, поэтому при исследовании каждого параметра стоит учитывать в большей степени не ***точность*** классификации, а разброс точности между анализом тестового и обучающего набора данных. Лучшие результаты ***общей точности*** были получены при следующих параметрах: число деревьев (*RF*:175, *XGB*:155), максимальное число признаков для разветвления в узле (*RF*:13), критерий примеси - индекс Джинни, максимальная глубина

(*RF*:19, *XGB*:10), минимальное число выборок в узле (*RF*:7, *XGB*:1), минимальное число выборок в листе (*RF*:1), доля признаков для обучения каждого дерева (*XGB*: 0,65), и доля выборки для обучения каждого дерева (*XGB*: 0,65). При этом ***точность*** *RF* увеличилась на 5%, а *XGB* – на 7%. Оценки работы классификатора показаны на Рисунке 5. Так, настройка модели *RF* повышает *Precision* до 12% (для *POPS*), а настройка *XGBoost* – *Recall* до 19% (для *POPS* и *SSL*).



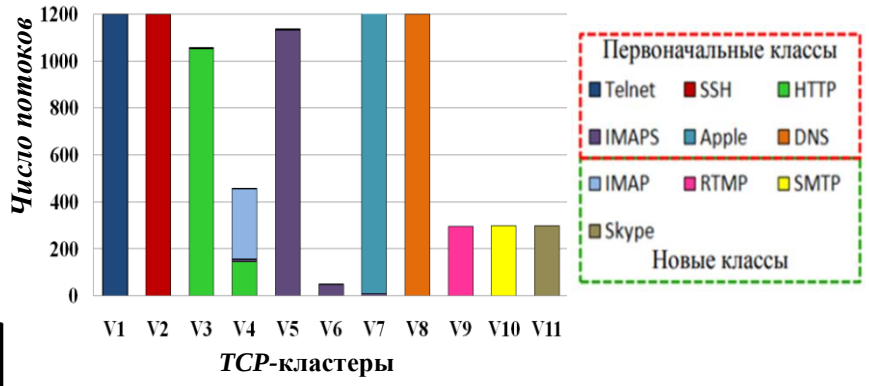

а) ***Общая точность*** б) Прирост ***F1-меры***

Рисунок 5. - Результаты классификации до и после расчета гиперпараметров

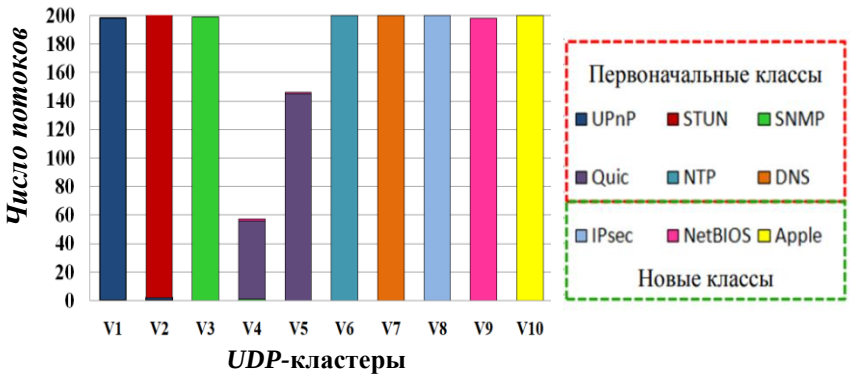
Основные результаты третьего раздела обсуждались на конференции *AIMEE2020* и опубликованы в статьях [4, 8].

**В четвертом разделе** модель классификации трафика расширяется за счет блоков кластеризации, которые позволяют ей добавлять новые классы в уже существующую модель (Рисунок 6).


а) Модель классификации




б) Результаты *TCP*-кластеризации




в) Результаты *UDP*-кластеризации

Рисунок 6. - Динамическая классификация трафика с возможностью добавления новых классов

В блоке сбора статистических данных (1) для поступающих в сеть первых 15 пакетов записываются время прихода и их размер. Далее, для каждого потока рассчитывается матрица признаков (2), в блоке предобработки данных (3) выполняется работа с выбросами и преобразование данных. В блоке классификации (4) с помощью *XGB* проводится классификация трафика по известным для модели классам. В том случае, если приложение является новым и незнакомым для модели, на этом этапе могут возникнуть некоторые ошибки, т.к. методы «обучения с учителем» способны классифицировать выборки только по известным для модели классам. Такое допущение сделано осознанно, т.к. и дальнейшая работа по управлению трафиком может быть основана только на известных классах. Тем не менее, блок классификации (4) представляет наиболее близкий класс для пришедшего нового потока и как следствие, для него применяются соответствующие механизмы управления сетью.

В то время как блок классификации (4) решает, к какому классу отнести активный поток, а контроллер применяет правила по управлению трафиком в режиме реального времени, копия информации о новом потоке из блока (3) отправляется в блок (6). Здесь хранится база данных по всем известным потокам. Каждая выборка представляется своими координатами в $N$-мерном пространстве в соответствии с определенными признаками. Эффективность кластеризации в данной работе повышается за счет введения блока (7), в котором проводится расчет расстояний между имеющимися выборками. В исследовании рассматривались пять способов построения матрицы расстояний: два из них являются стандартными и общепринятыми методами (расстояние *Евклида* и *Манхэттена*), а три других основаны на методах предрасчитанных расстояний (*Random Forest*, *Extremely Randomized Trees - ET* и *Random Trees Embedding - RTE*), общий алгоритм которых представляется следующим образом:

1. Расчет матрицы признаков $n \times m$ на основе статистических характеристик потока, где $n$ – количество потоков, а $m$ – количество признаков.

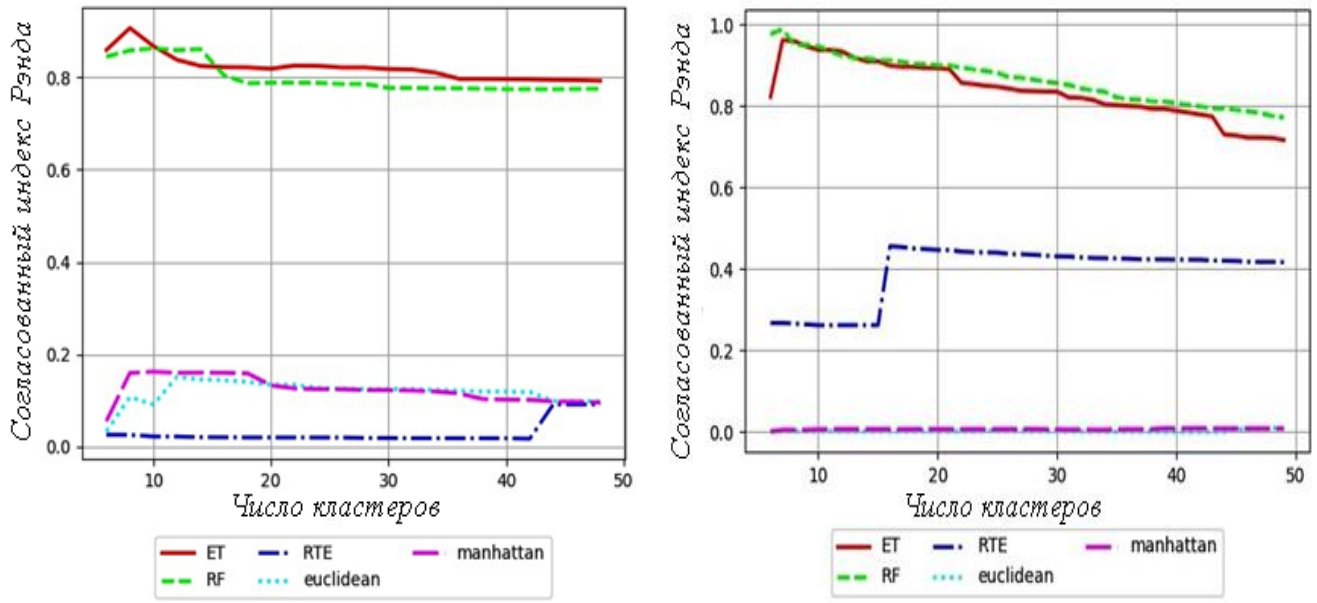
2. Построение леса из $k$ деревьев на основе матрицы признаков обучающей последовательности, для *ET* и *RF* проходит в контролируемом режиме.

3. Присвоение каждому потоку индекса листа, на котором он оказался на каждом из деревьев. В результате этого шага получается матрица листьев $n \times k$.

4. Создание матрицы попарной близости $n \times n$, где между каждой парой потоков указывается общее количество листьев, на которых они оказались вместе, среди всех деревьев. Процесс сравнения листьев проводится с помощью алгоритма *One Hot Encoding* (представление каждого состояния с помощью одного триггера) и произведения закодированной матрицы индексов и ее же в транспонированном виде.

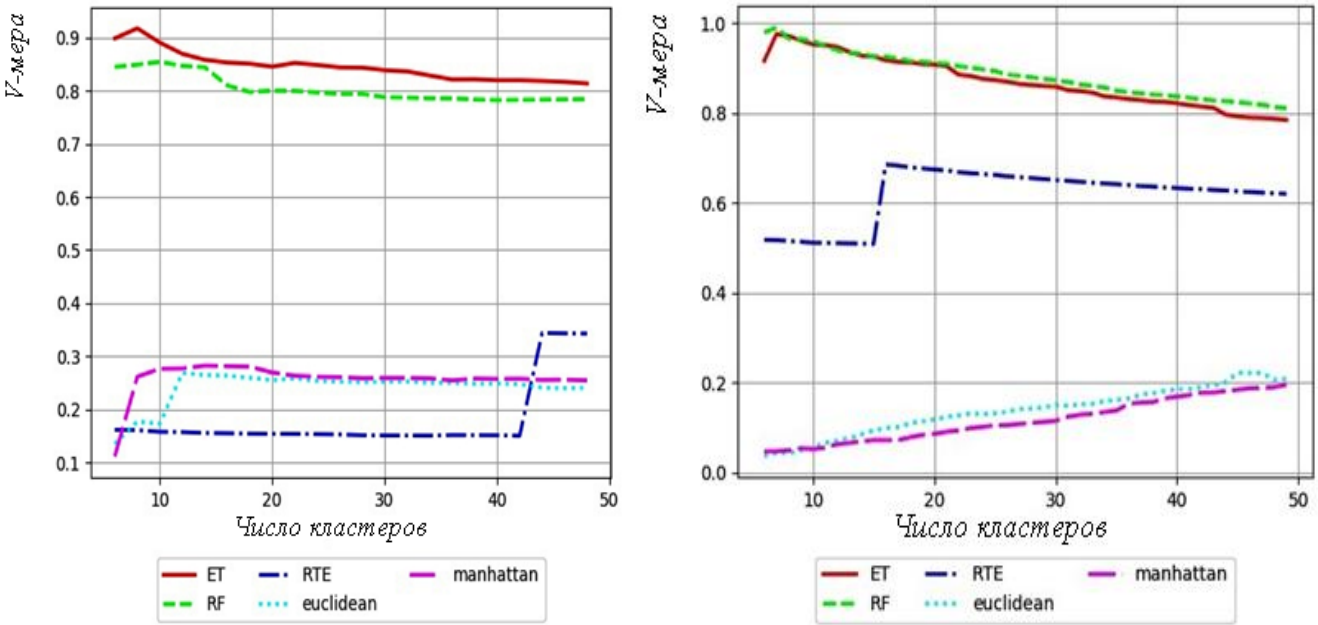
5. Получение матрицы расстояний путем нормирования матрицы близости относительно максимального значения и вычитания ее значений из единицы.

Результаты кластеризации трафика при применении разных матриц расстояний (Рисунок 7-8) показывают неприемлемость методов бесконтрольного расчета расстояний на основе расстояний *Евклида, Манхэттена* и *RTE* и высокую эффективность методов контролируемого построения расстояний на основе *RF* и *ET*.




а) *TCP* – трафик б) *UDP*-трафик

Рисунок 7. - Результаты кластеризации: ***согласованный индекс Рэнда*** при изменении числа кластеров от 5 до 45 для разных матриц расстояний




а) *TCP* – трафик б) *UDP*-трафик

Рисунок 8. - Результаты кластеризации: ***V-мера*** при изменении числа кластеров от 5 до 45 для разных матриц расстояний

После построения матрицы расстояний между выборками, в блоке кластеризации (8) проводится разделение выброк на кластеры с помощью агломеративной кластеризации, которая позволяет регулировать количество кластеров в зависимости от расстояния между кластерами, т.е. в отличие от других распространенных методов, таких как *K*-средних, спектральная кластеризация и т.д., не требует информации о числе кластеров до кластеризации. Эта особенность позволяет не только определить оптимальное число кластеров для обучающей выборки, но, что более важно, появляется возможность вводить новые классы в уже существующую модель автоматически, регулируя лишь расстояние между кластерами.

Для исследования блока кластеризации трафик был разделен на *DNS, Apple, IMAPS, HTTP, SSH, Telnet* (1-я группа) и *RTMP, IMAP, SMTP, Skype* (2-я группа). Первая группа разбита на кластеры, так, чтобы *ARI* достигал высоких значений (0,9-1,0), и в то же время кластеры не содержали бы в себе малое количество выборок (по 1-2). Так было получено 8 кластеров *TCP*: 1 *DNS*, 1 *Telnet*, 1 *Apple*, 2 *IMAPS*, 1 *SSH*, 2 *HTTP* и 7 кластеров *UDP*: 1 *UPnP*, 1 *STUN*, 1 *SNMP*, 2 *Quic*, 1 *NTP*, 1 *DNS* с минимальной относительной дистанцией между кластерами 0,9997. Далее поочередно добавляются приложения из второй группы, при этом, если итоговое число кластеров оказывается меньше исходного, минимальная дистанция для разделения на кластеры уменьшается, что позволяет избежать объединения между собой соседних кластеров за счет появления промежуточных классов. В эксперименте удалось обнаружить, что при добавлении каждого нового класса добавляется новый кластер, кроме случая с добавлением *IMAP*. Это объясняется наличием небольшого кластера *HTTP*, который объединился с выборками *IMAP*. Аналогичный эксперимент проводился и для *UDP*, результаты представлены на Рисунке 6 (б, в).

Для работы модели требуется около 200-250 Мбит ОЗУ (для *Intel Core Processor, Haswell, no TSX, IBRS 2394.454 MHz CPU*), *Python 3.7*, режим обновления модели занимает около 0,1-0,12 мс, а режим классификации ≈ 2 мкс (при одновременной классификации до 100 потоков). Выбор частоты обновления модели основывается на динамике изменения приложений сети и требованиям к качеству обслуживания с учетом технических характеристик оборудования.

Результаты четвертого раздела обсуждались на конференциях *CSDEIS2020* и *28-й FRUCT* (2021) и опубликованы в статьях [3; 5].

**В пятом разделе** разрабатывается алгоритм сбора статистической информации для матрицы признаков режима реального времени. Алгоритм представлен для *SDN*-сетей в *P4*-коммутаторах (*Programming Protocol – independent Packet Processors*), которые позволяют запрограммировать процесс обработки пакета в коммутаторе, благодаря чему сетевые элементы могут работать эффективнее, снижается задержка передачи пакета, и рационально используются ресурсы сети (Рисунок 9).

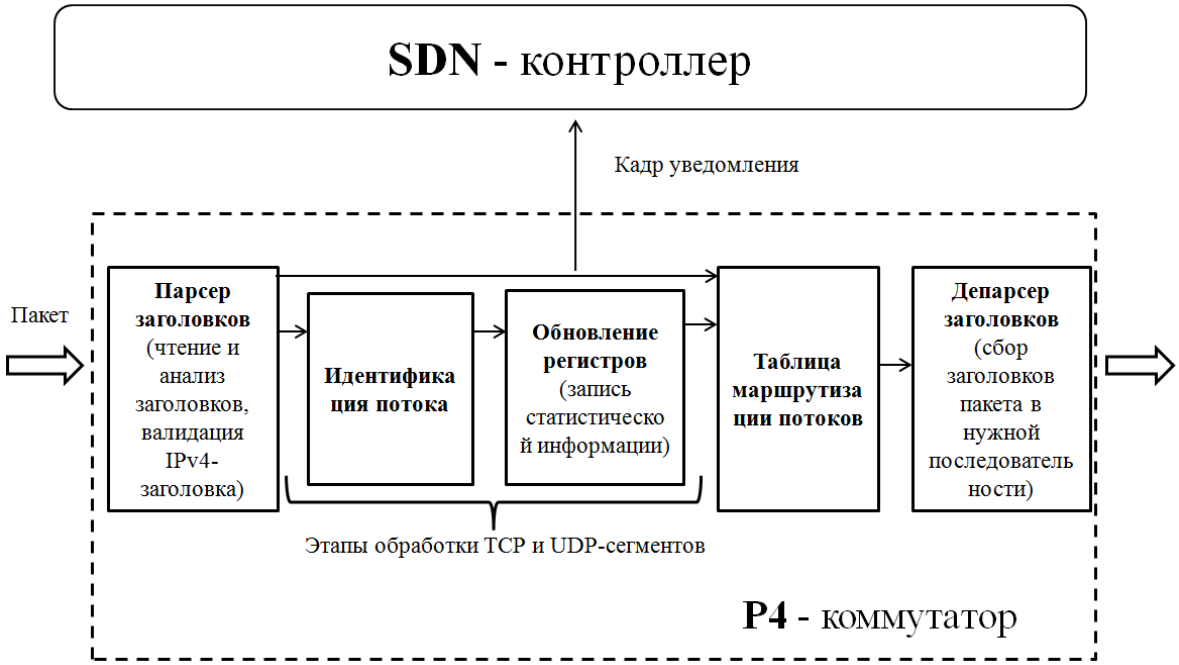



Рисунок 9. – Алгоритм сбора статистической информации о пакетах

На этапе парсера заголовков проводится чтение и анализ заголовков, а также валидация *IPv4*-заголовка. Следующие два блока предназначены только для *TCP* и *UDP* – сегментов (при необходимости, структура языка и создаваемых на нем коммутаторов *P4* позволяет легко вводить любые необходимые заголовки, даже неопределенные ни в какой спецификации). На этапе идентификации потоков определяется, к какому из имеющихся потоков относится пришедший пакет, при необходимости создается новый поток. Далее следует обновление регистров, при котором добавляется информация о новом пришедшем пакете. Когда из одного *TCP* - потока приходит 15-й по счету пакет (или 10-й для *UDP*), вся информация отправляется в *SDN*-контроллер. Таблица маршрутизации позволяет проводить маршрутизацию потоков и построена аналогично *Flow_Table OpenFlow* - коммутаторов, но с более расширенными возможностями. На этапе депарсера заголовки собираются в нужной последовательности.

Блоки идентификации потоков и обновления регистров разрабатывались на основе специальных ячеек памяти, называемых регистрами. В предложенной реализации коммутатора, исходя из количества и предназначений ячеек памяти, можно выделить три типа регистров: одиночный регистр, регистры потоков и регистры пакетов (Рисунок 10).

Одиночный регистр носит название *flow_ID* и представляет собой только одну ячейку памяти. Он служит уникальным идентификатором потоков внутри коммутатора. Значения, которые в нем встречаются – это адреса ячеек памяти регистров, в которых записывается информация о соответствующем потоке. То значение, которое имеется во *flow_ID* в выбранный момент времени, показывает номера ячеек памяти, в которые будет записан следующий новый поток.

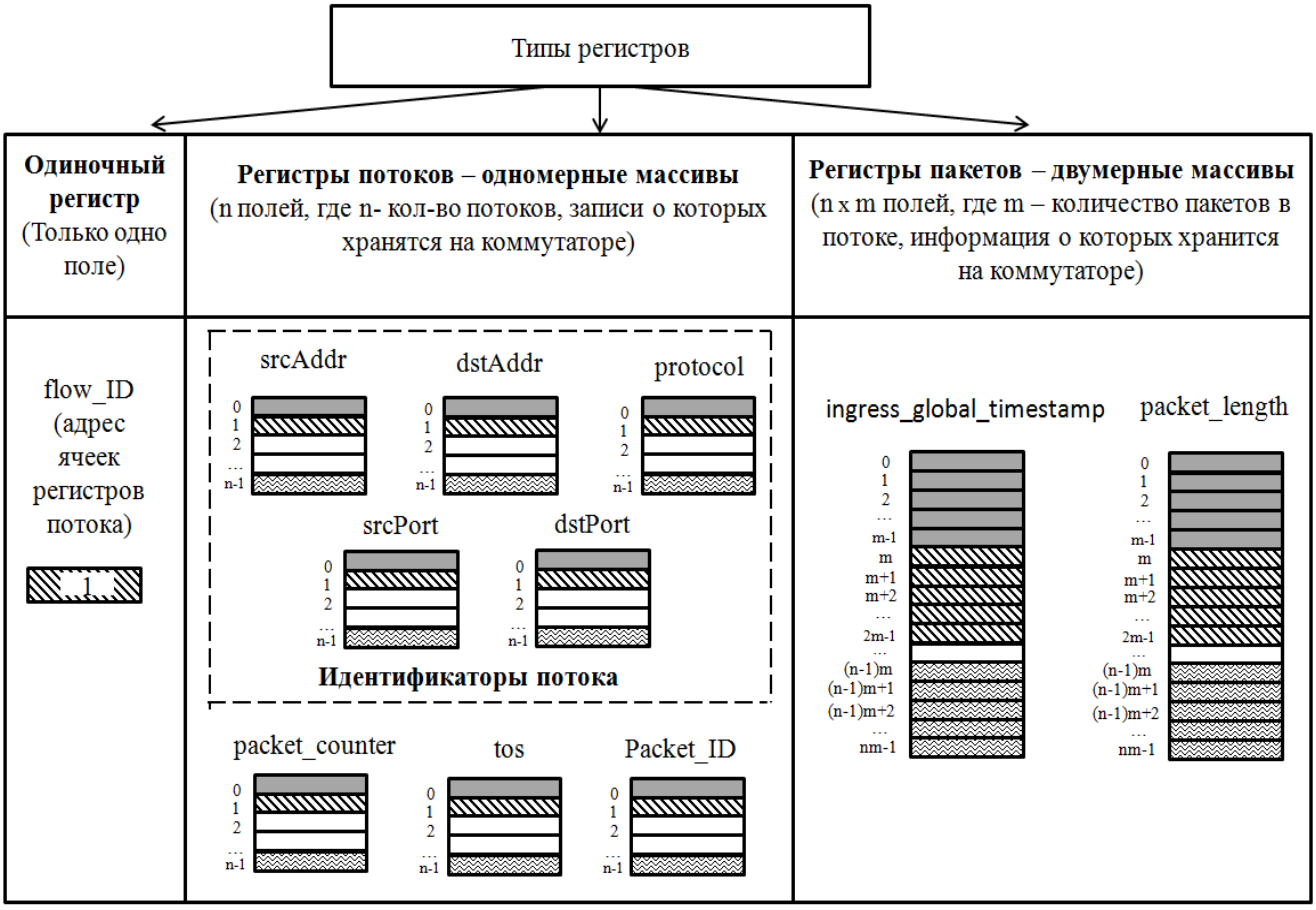



Рисунок 10. - Система организации памяти в *P4*-коммутаторе

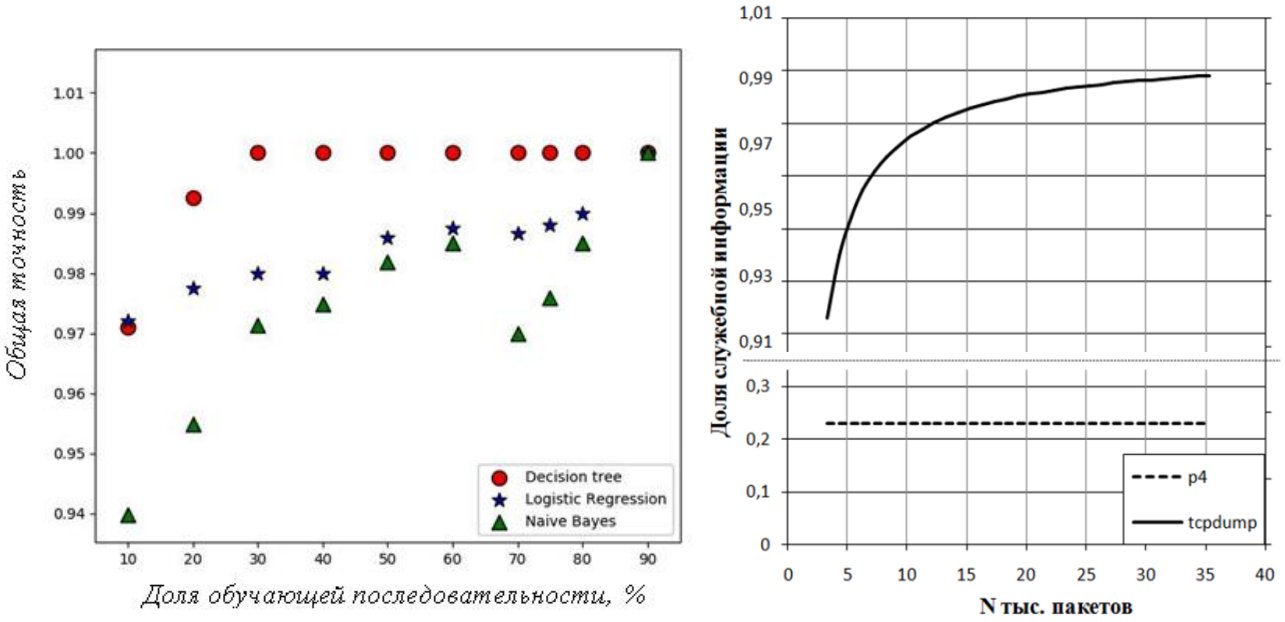
Регистры потоков – это одномерные массивы, размер которых *n*-совпадает с числом потоков, информация о которых должна храниться в памяти коммутатора. К регистрам потоков относятся: *srcAddr* (*IP*-адрес источника), *dstAddr* (*IP*-адрес назначения), *protocol* (протокол транспортного уровня), *srcPort* (порт источника), *dstPort* (порт назначения), *tos* (значение поле *DSCP IP* – пакета); *packet_counter* (счетчик количества пакетов потока, прошедших через коммутатор), *Packet_ID* (аналог *flow_ID* для пакетов).

Поля регистров *srcAddr* (4 байта), *dstAddr* (4 байта), *protocol* (1 байт), *srcPort* (2 байта), *dstPort* (2 байта) вместе образуют значение *5-tuple*, которое используется для идентификации потока на уровне всей сети. Поля регистра *tos* (1 байт) могут применяться не только непосредственно для обеспечения *QoS* стандартными методами, но и в качестве возможного маркера с целью дальнейшей классификации потоков методами машинного «обучения с учителем».

Регистры пакетов представляют собой двумерные массивы, организованные из одномерных, адреса ячеек в которых определяются с помощью регистров *flow_ID* и *Packet_ID*. Для целей классификации трафика методами машинного обучения были введены два вида регистров: *ingress_global_timestamp* (14 байт) как время прихода пакета на этап обработки коммутатора, выраженное в мкс и *packet_length* (10 байт) – полная длина пакета в байтах. Размер

такого регистра $n \times m$ полей, где $n$ – число потоков, а $m$ –число пакетов, информацию о которых следует хранить в памяти коммутатора. Например, информация о 2-м пакете 0-го потока хранится в ячейках памяти с номером 0 в регистрах потоков (*srcAddr*[0], *dstAddr*[0] и т.д.) и в ячейках с номером 2 в регистрах пакетов (*packet_length*[2], *ingress_global_timestamp*[2]).

Для оценки работы модели был поставлен натурный эксперимент в виртуальной сети *Mininet*. Генератором трафика выступал инструмент *D-ITG* (*Distributed Internet Traffic Generator*), который создал 1000 различных *UDP*-потоков, поступающих в сеть в случайные моменты времени по экспоненциальному закону. Три приложения имитировали работу онлайн-игр: 195 потоков *Quake3*, 187 потоков *CSa* (активный режим игры *Counter-Strike*) и 191 поток *CSi* (неактивный режим *Counter-Strike*) и два приложения – передачу голосового трафика (кодеки *G.711.1* – 207 потоков и *G.729.2* – 216 потоков). Результаты классификации (Рисунок 11) подтверждают эффективность разработанного алгоритма сбора статистической информации и способность модели классификации работать не только с различными типами приложений, но и с различными режимами одного и того же сервиса (*CSa* и *CSi*) при наличии соответствующей тренировочной базы.




а) Точность классификации б) Доля служебной информации

Рисунок 11. - Результаты классификации трафика на *P4*-коммутаторе

Для сравнительной оценки работы разработанного метода сбора статистической информации через программирование *P4*-сетей и других популярных инструментов мониторинга сетей (рассматривается на примере *tcpdump*) предлагается оценить объемы информации и долю служебной, неиспользуемой в матрице признаков информации.

Для разработанного метода передается *5-tuple, tos* и значения регистров *packet_length* и *ingress_global_timestamp* для первых $n$ пакетов потока. Если принять $n$=15, а размер заголовка

пакета, в котором передается служебная информация, принять равным 32, то для одного потока передается 196 байт информации, для *k* потоков, длиной более 15 пакетов, $V_{инф\ p4}$=196*k*. Тогда доля служебной информации при передаче результатов сбора статистической информации контроллеру, постоянна и равна $\alpha_{p4} = 0{,}23$.

При использовании *tcpdump 5-tuple* и *tos* будет передаваться для каждого пакета. Более того, собирается информация о потоках, длиной менее 15 пакетов и информация о пакетах, начиная с 16-го. Таким образом, объем информации, передаваемой с помощью *tcpdump*: $V_{инф\ tcpdump}$=54*S*, где *N*-количество всех пакетов всех потоков. Пусть $\{n_1, n_2 \dots, n_k\}$ – множество, в котором каждый элемент $n_i \geq 15$ - количество пакетов в *i*-м потоке, а в множестве $\{m_1, m_2 \dots, m_h\}$ каждый элемент $m_j < 15$ для *h* потоков. Тогда доля служебной информации, передаваемая в этом случае, приблизительно оценивается как:

$$\alpha_{tcpdump} = 1 - \frac{75k}{28\left(\sum_{i=1}^{k} n_i + \sum_{j=1}^{h} m_j\right)}, \quad (10)$$

а разность объема передаваемой информации при использовании инструментов *tcpdump* и *P4* (байт):

$$\Delta V_{\text{инф}} = V_{\text{инф } tcpdump} - V_{\text{инф } p4} = 56\left(\sum_{i=1}^{k} n_i + \sum_{j=1}^{h} m_j\right) - 196k. \quad (11)$$

На Рисунке 11 (б) изображен сравнительный график зависимости доли служебной информации от *N* - количества пакетов во всех потоках для *k*=100 потоков. Видно, что доля служебной информации для передачи данных при использовании стандартных способов мониторинга находится в пределах от 0,91 до 1 и стремится к 1 при увеличении числа пакетов в сети. Таким образом, разработанный метод сбора статистической информации позволяет значительно снизить объемы служебной информации в сети.

Результаты пятого раздела представлялись на конференциях *CSDEIS2019* и *ИТММС 2019* и были опубликованы в статьях [1; 9].

## ЗАКЛЮЧЕНИЕ

Основные результаты диссертационной работы сводятся к следующему:

1. Обосновано применение методов машинного обучения для решения задачи динамической классификации трафика с целью обеспечения качества обслуживания в мультисервисных *SDN*-сетях.

2. Разработана матрица признаков для классификации трафика с целью обеспечения *QoS*, в которой признаками являются индивидуальные статистические параметры первых 10-15 пакетов, такие как длина и межинтервальное время прихода пакета на интерфейс. Такой подход

имеет два значительных отличия от общепринятых методов построения матрицы признаков. Во-первых, он не требует никакой информации о *TCP*-флагах и заголовках вышележащего уровня, что делает разработанную матрицу признаков инвариантной по отношению к типам потоков трафика. Во-вторых, набор параметров, основанный только на первых 10-15 пакетах, позволяет эффективно применять классификацию к активным потокам в режиме реального времени. Точность полученного классификатора оказалась на 10-25% выше результатов на основе других матриц признаков.

3. Разработана и исследована статическая модель классификации трафика методами машинного «обучения с учителем» на основе созданной матрицы признаков, отличающаяся от аналогичных содержанием структурных блоков и их расположением. На каждом этапе работы модели проводятся процедуры, повышающие эффективность классификации для соответствующих методов машинного обучения.

Прирост точности классификации для метода *XGBoost* в блоке предварительной обработки данных (квантильная трансформация и удаление выбросов) – 6,5-7,2%, в блоке классификатора за счет настройки гиперпараметров — 18-20% , а суммарно до 26%, и в целом точность классификации составила значение в 91%, что на 3% выше, чем точность классификации методом *Random Forest*. Эти результаты показывают перспективность применения метода *XGBoost*, что позволяет расширить существующий набор методов классификации потоков трафика для задач поддержания *QoS*.

4. Разработана модель эффективной кластеризации трафика (*ARI* около 0,9-1,0), адаптированная к сформированной матрице признаков, за счет применения в процессе кластеризации матрицы расстояний, предрассчитанной на основе результатов классификации потоков методами *Extremely Randomized Trees* и *Random Forest*. Показано также, что наиболее распространенные и стандартные подходы к расчету матриц расстояний, такие как расстояния *Евклида* и *Манхэттена*, оказались непригодными к применению в таких условиях.

5. Разработан алгоритм гибкого сбора статистической информации о пакете в *P4*-коммутаторах, основанный на разработанной системе организации памяти, позволяющий хранить и передавать на контроллер статистическую и идентификационную информацию о любом пакете и только по мере необходимости (например, после накопления первых 10-15 пакетов), что снижает служебную нагрузку на сеть и устройства по сравнению с традиционными вариантами полного мониторинга сетевых элементов в 4-4,5 раза.

6. Создана не имеющая аналогов в открытых исследованиях модель классификатора трафика, полученная за счет объединения точности и скорости работы методов «обучения с учителем» для работы в режиме реального времени с возможностью добавления новых классов за счет методов «обучения без учителя» и уточнения существующих кластеров. Несмотря на то, что

работа ориентирована на задачи *SDN*-сети, результаты работы могут быть использованы и в других сетях.

В дальнейших исследованиях планируется разработка методов управления трафиком на основе полученных классов для оптимального использования сетевых ресурсов и обеспечения качества обслуживания потоков поступающего трафика.

## СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

### Статьи, опубликованные в научных изданиях, индексированных в международных наукометрических базах, в т.ч. WoS, Scopus и Springer

1. **Krasnova, I.A.** Collection of Individual Packet Statistical Information in a Flow Based on P4-switch / **I.A. Krasnova**, V.Yu. Deart, V.A. Mankov // In: Hu Z., Petoukhov S., He M. (eds) Advances in Intelligent Systems, Computer Science and Digital Economics. CSDEIS 2019. / Advances in Intelligent Systems and Computing; Springer, Cham - 2020. - vol 1127, - pp.106-117, doi:10.1007/978-3-030-39216-1_11

2. **Krasnova, I.A.** Development of a Feature Matrix for Classifying Network Traffic in SDN in Real-Time Based on Machine Learning Algorithms / **I.A. Krasnova**, V.Yu. Deart, V.A. Mankov // 2020 International Scientific and Technical Conference Modern Computer Network Technologies (MoNeTeC) - Moscow, 2020. - pp. 1-9. - doi:10.1109/MoNeTeC49726.2020.9258314

3. **Krasnova, I.A.** Agglomerative Clustering of Network Traffic Based on Various Approaches to Determining the Distance Matrix / **I.A. Krasnova**, V.Yu. Deart, V.A. Mankov // 2021 28th Conference of Open Innovations Association (FRUCT) - Moscow, 2021. - pp. 81-88. - doi:10.23919/FRUCT50888.2021.9347616

4. **Krasnova, I.A.** Evaluation of the Effect of Preprocessing Data on Network Traffic Classifier Based on ML Methods for QoS Predication in Real-Time / **I.A. Krasnova**, V.Yu. Deart, V.A. Mankov // In: Hu Z., Petoukhov S., He M. (eds) Advances in Artificial Systems for Medicine and Education IV. AIMEE 2020. / Advances in Intelligent Systems and Computing, Springer, Cham. -2021. - vol 1315, - pp. 55-64. - doi:10.1007/978-3-030-67133-4_5

5. **Krasnova, I.A.** An Extensible Network Traffic Classifier Based on Machine Learning Methods / **I.A. Krasnova**, V.Yu. Deart, V.A. Mankov // In: Hu Z., Petoukhov S., He M. (eds) Advances in Intelligent Systems, Computer Science and Digital Economics II. CSDEIS 2020. / Advances in Intelligent Systems and Computing, Springer, Cham. -2021. - vol 1402, - pp. 10-19. – doi: 10.1007/978-3-030-80478-7_2

**Статьи в ведущих рецензируемых научных журналах и изданиях, рекомендованных ВАК**

6. **Краснова, И.А.** Алгоритм динамической классификации потоков в мультисервисной SDN-сети / **И.А. Краснова**, В.А. Маньков // T-Comm: Телекоммуникации и транспорт. - 2017. – Т.11, №12. - С. 37-42.

7. **Краснова, И.А.** Анализ перспективных подходов и исследований по классификации потоков трафика для поддержания QoS методами ML в SDN-сетях / **И.А. Краснова**, В.Ю. Деарт, В.А. Маньков // Вестник СибГУТИ - 2021. - №1. - С. 3-22.

8. **Краснова, И.А.** Анализ влияния параметров алгоритмов Machine Learning на результаты классификации трафика в режиме реального времени / **И.А. Краснова** // T-Comm: Телекоммуникации и транспорт. - 2021. – Т.16, №9. - С. 24-35.- doi: 10.36724/2072-8735-2021-15-9-24-35

**Доклады на научных конференциях**

9. **Краснова, И.А.** Классификация потоков трафика SDN-сетей методами машинного обучения в режиме реального времени / **И.А. Краснова**, В.А. Маньков // Труды международной научно-технической конференции «Информационные технологии и математическое моделирование систем 2019». – Одинцово, 2019. - С.65-68.- doi:10.36581/CITP.2019.31.51.016

10. **Краснова, И.А.** Задача управления трафиком с динамическим определением QoS в мультисервисных SDN сетях / **И.А. Краснова**, В.А. Маньков // Сборник трудов XI международной отраслевой научно-технической конференции «Технологии информационного общества» / МТУСИ. Москва, 2017. - С.67-68.

11. **Краснова, И.А.** Исследование параметров качества сети передачи данных с помощью моделирования в среде ns-3 / **И.А. Краснова** // Десятая московская научно-практическая конференция: «Студенческая наука», сборник тезисов / Московский студенческий центр. – Москва, 2015. –Т.3. - С.691-693.

**Учебное пособие**

12. **Краснова, И.А.** Виртуализация сетевых функций и программно-конфигурируемые сети: учебное пособие / **И.А. Краснова**, В.А. Маньков, А.Е. Панов; МТУСИ – Москва, 2020.– 126 с.