На правах рукописи

Ванюшина Анна Вячеславовна

Классификация IP-трафика в компьютерной сети с использованием алгоритмов машинного обучения

Специальность 05.13.15 –

Вычислительные машины, комплексы и компьютерные сети

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Москва — 2019

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

Научный руководитель:	Ерохин Сергей Дмитриевич, кандидат технических наук, доцент, ректор МТУСИ
Официальные оппоненты:	Басараб Михаил Алексеевич, доктор физико-математических наук, заведующий кафедрой ИУ8 МГТУ им. Н.Э. Баумана
	Смычѐк Михаил Александрович, кандидат технических наук, доцент, заместитель начальника отдела АО «Гипрогазцентр»
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В.А. Трапезникова Российской академии наук

Защита состоится 13 февраля 2020 г. в 13 час. 00 мин. на заседании диссертационного совета Д 212.131.05 по адресу: РТУ МИРЭА, Москва, проспект Вернадского, д. 78, ауд Д 117.

С диссертацией можно ознакомиться в библиотеке и на сайте РТУ МИРЭА https://www.mirea.ru/

Автореферат разослан «11» января 2020 г.

Учёный секретарь
диссертационного совета Д 212.131.05,
кандидат технических наук Е.Г.Андрианова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Классификация IP-трафика является важной задачей для управления трафиком, улучшения технико-экономических, эксплуатационных характеристик и защиты компьютерных сетей (КС), поскольку позволяет определить тип и структуру приложения, которое является его источником. Системы классификации сетевого трафика используются в широком спектре сетевых функций и приложений: обеспечение качества связи (QoS, QoE), выполнение политик информационной безопасности, в том числе при разработке алгоритмов и программного обеспечения, обеспечивающих контроль, диагностику состояния КС и выявление сетевых проблем, сбор статистических данных и т.д.

Традиционные методы классификация сетевого трафика (например, классификация по номерам портов протоколов транспортного уровня) обладают рядом существенных недостатков, что является причиной роста исследований в этом направлении. Существенный рост объемов и видов сетевых протоколов за последние годы повышают актуальность проблемы разработки методик и алгоритмов классификации трафика с пониженной вычислительной сложностью. Особо остро стоит задача классификации трафика с использованием протоколов шифрования.

Одним из наиболее перспективных направлений классификации сетевого трафика являются статистические методы, основанные на выявлении и анализе статистических характеристик трафика. Наиболее перспективными здесь являются технологии машинного обучения (МО, англ.ML, Machine Learning) и интеллектуального анализа данных (ИАД, англ.DM, Data Mining), получившие широкое распространение в смежных областях науки.

Степень разработанности темы диссертационного исследования

В диссертационном исследовании решается задача обучения по прецедентам – классификация объектов на основе известной совокупности их признаков (атрибутов), с целью совершенствования теоретической и технической базы компьютерных сетей, обладающих высокими качественными и эксплуатационными показателями, на примере использование стека TCP/IP. Задача заключается в отнесении объекта к одному из заранее заданных, непересекающихся классов, по правилам которые извлекаются из обучающей выборки, содержащей аналогичные, но уже классифицированные объекты.

Теоретическую базу диссертации в области методов МО и ИАД составили работы таких ученых, как Айвазян С.А., Айзерман М.А., Барсегян А.А., Загоруйко Н.Г., Вапник В.Н., Воронцов К.В., Mitchell T., Hastie T., Tibshirani R., Friedman J.T, Xu K., Zhang Z., Bhattacharyya S., Heckerman J.D. и др., выполнен-

ных преимущественно в области экономических исследований, которые, как оказалось, можно встроить в предмет сетевых и телекоммуникационных исследований.

Отдельные вопросы построения и исследования классификации трафика с помощью методов МО рассматривались в трудах Большева А.К., Гетьмана А.И., Зубкова Е.В., Котенко И.В., Козьмовского Д.В., Михайлова А.Ю., Маркина Ю.В., Назарова А.О., Петровского М.И., Санарова А.С., Шелухина О.И., Щербаковой Н.Г., M.Pietrzyk, Z.Chen, B.Yang, J.Erman, K. Balachandran, J.H.Broberg, T.Bujlow, V.Carela-Español, C.C. Aggarwal, Y.Wang, J.Erman, M.Arlitt, A.Mahanti и др.

Однако в работах этих авторов не получили достаточного отражения теоретические и практические вопросы классификации приложений в КС, использующих стек протоколов TCP/IP в условиях априорной неопределенности (наличии «фонового» трафика), а также оценки эффективности различных алгоритмов, реализующих методы МО при потоковом режиме поступления данных. Вышесказанное обусловливает научную актуальность настоящего исследования в области эффективной классификации IP-трафика на основе методов машинного обучения.

Целью диссертационного исследования является повышение эффективности классификации приложений сетевого трафика компьютерных сетей в условиях априорной неопределенности методами машинного обучения.

Решаемые задачи:

Для достижения цели в настоящей работе решены следующие задачи:

1. Экспериментальные исследования функционирования КС с целью обоснования рациональной структуры и объема обучающей и тестирующей выборок анализируемых сетевых приложений, и оценка их влияния на эффективность классификации методами МО.
2. Повышение защищенности КС, использующих телекоммуникационные технологии, анализ полученных численных оценок эффективности алгоритмов классификации сетевых приложений в условиях априорной неопределенности структуры анализируемых данных при потоковом режиме их поступления.
3. Разработка и реализация алгоритма обнаружения смены концепта при потоковом режиме поступления данных на основе текущего анализа статистических характеристик анализируемых приложений в КС, использующих стек TCP/IP.
4. Разработка специального программного обеспечения для обеспечения контроля и диагностики функционирования КС путем автоматической классификации анализируемых приложений методом машинного обучения .

Новизна научных результатов

1. Впервые показано, что для сохранения высокой эффективности обработки анализируемых приложений в КС, использующих стек протоколов TCP/IP (FTP, Web, Mail, SSH, Skype) можно до 50% сократить число анализируемых признаков (атрибутов), что незначительно (не более чем на 5…7%) снижает достоверность классификации.

2. Показано, что путем применения специальной обработки приложений с использованием МО можно улучшить взаимодействие и защиту КС. При этом для обеспечения высокой достоверности (не ниже 95%) при большом объеме измерений и уменьшенном числе необходимых атрибутов наилучшим образом подходят алгоритмы классификации C4.5 и Random Forest (RF) .

3. Для преодоления проблемы априорной неопределенности анализируемых приложений в КС, построенных с использованием телекоммуникационных технологий, введен в рассмотрение дополнительный класс «Неизвестное приложение», показавший результаты, незначительно уступающие по качественным показателям результатам классификации «с учителем».

4. Для контроля за изменением текущих статистических характеристик анализируемых атрибутов приложений в КС, построенных с использованием телекоммуникационных технологий, разработан новый алгоритм обнаружения смещения концепта наблюдаемого потока данных с помощью двух скользящих окон.

Теоретическая и практическая значимость диссертационной работы

Теоретическая значимость работы состоит в:

- численных результатах классификации приложений по потокам и пакетам в КС, использующих стек протоколов TCP/IP, методами машинного обучения;

- научно-обоснованных рекомендациях по выбору параметров исследуемых алгоритмов в условиях априорной неопределенности, в том числе в режиме реального времени;

- предложенном методе классификации приложений в КС, использующих стек протоколов TCP/IP, с дополнительным классом «Неизвестное приложение», показавшем высокую эффективность по сравнению с обычно используемыми методами, что позволит повысить защищенность КС в условиях априорной неопределенности методами МО;

- новом алгоритме обнаружения «смены концепта» наблюдаемого потока с помощью двух скользящих окон, контролирующих изменение текущих статистических характеристик атрибутов анализируемых приложений КС;

Практическая значимость работы заключается в:

- в возможности реализации разработанных алгоритмов классификации и

обнаружения смены концепта в потоковом режиме в задачах обработки приложений в КС, построенных с использованием различных телекоммуникационных технологий;

- разработке специального программного обеспечения (СПО) для обеспечения контроля и диагностики функционирования КС, использующих стек протоколов TCP/IP путем автоматической классификации приложений, в том числе в реальном масштабе времени;

- использовании результатов работы в аппаратно-программных комплексах в ТБ КЦ ПАО «МТС», ООО «Эльстер Метроника» и учебном процессе МТУСИ.

Методология и методы исследования. В работе используются методы исследования теории вероятности, математической статистики, математического имитационного моделирования, машинного обучения и интеллектуального анализа (обработки) данных.

Положения, выносимые на защиту

1. Экспериментальные результаты исследования функционирования компьютерных сетей, рациональную структуру, объем обучающей и тестирующей выборок анализируемых приложений, оценки их влияния на эффективность классификации с помощью алгоритмов МО, показавшие, что для достижения точности не менее 99,5% достаточно обработать не более 35 последовательных пакетов.

2. Количественный состав и структура атрибутов, необходимых для эффективной классификации приложений в КС, использующих стек протоколов TCP/IP, позволившие сократить их число на 33%, при незначительном (не более чем на 5…7%) снижении достоверности правильной классификации, но значительном упрощении процесса обработки анализируемых данных.

3. Способ преодоления проблемы априорной неопределенности анализируемых приложений в КС, построенных с использованием телекоммуникационных технологий, путем введения в рассмотрение дополнительного класса «Неизвестное приложение», показавшего, что при его использовании снижение вероятности ложной классификации может достигать более 30% при снижении достоверности правильной классификации не более 2,5%, что значительно превосходит эффективность алгоритмов кластеризации k-Means и DBSCAN.

4. Алгоритм обнаружения «смены концепта» наблюдаемого потока данных в КС, построенных с использованием телекоммуникационных технологий, указывающий на необходимость обновления текущей модели классификации и отличающийся от известных более высоким быстродействием

за счет учета статистических характеристик анализируемых атрибутов приложений.

Достоверность результатов исследования подтверждается: корректным использованием современного математического аппарата; достаточно широкой апробацией результатов, подтверждением адекватности моделей численными экспериментами на базе долговременной выборки реального IP-трафика ЦОД.

Апробация результатов работы

Основные результаты исследования были представлены и получили положительную оценку на Международном форуме информатизации (Москва, 2014-2017 гг.), международной научно-технической конференции «Телекоммуникационные и вычислительные системы» (Москва, 2017г, 2018.), на отраслевой научно-технической конференции «Технологии информационного общества» (Москва, 2007-15 гг.2017-2018гг.), XXII Международной научной конференции (WECONF-2019г.).

Разработанное автором программное обеспечение для решения задачи классификации методами машинного обучения в реальном масштабе времени внедрено в ТБ КЦ ПАО «МТС» и ООО «Эльстер Метроника», а полученные теоретические результаты использованы в учебном процессе МТУСИ, что подтверждено соответствующими актами.

Публикации по теме диссертации включают 13 научных публикаций, в том числе 8 статей общим объемом 3,51 п.л., 7 из которых опубликованы в рецензируемых периодических изданиях перечня ВАК РФ при Минобрнауке РФ.

Вклад соавторов заключался в формулировке задач и обсуждении полученных результатов.

Личный вклад автора. Все основные научные результаты диссертации получены автором лично. Работа представляет собой законченное, доведенное до инженерного уровня решение актуальной научно-технической задачи.

Объем и структура работы. Диссертация состоит из введения, 4-х глав, общих выводов, заключения, списка сокращений и списка литературы, содержащего 121 наименование. Основное содержание работы изложено на 147 страницах машинописного текста. Имеется 21таблица, 67 рисунков и 4 Приложения, общий объем диссертации 201 страница.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** сформулированы объект, предмет, цель, научная и частные задачи и границы исследования, показана актуальность работы, сформулированы основные положения, выносимые на защиту.

В **первом разделе «Задачи и проблемы классификации приложений КС использующих стек протоколов TCP/IP методами машинного**

обучения» анализируются особенности классификации приложений КС. Наиболее простым решением для классификации приложений является выявление и использование номеров портов транспортных протоколов. Данный метод имеет много преимуществ: он очень «быстрый», потребляет мало ресурсов, его поддержка реализована во многих сетевых устройствах. Он не анализирует полезную нагрузку приложения, поэтому не подвергает опасности конфиденциальность пользователей. Однако метод полезен только для выявления протоколов и приложений, использующих фиксированные номера портов. Кроме того, некоторые приложения могут использовать известные номера портов, чтобы «обмануть» политику безопасности в конкретной сети.

Как показывает анализ, статистические методы являются оптимальным выбором для классификации приложений КС в реальном времени, обеспечивая приемлемое потребление ресурсов при заданной высокой достоверности. Обеспечение контролируемости проверяемых данных и соблюдение законодательства страны можно обеспечить путем запрета глубокого анализа пакетов в сети, поскольку могут возникать проблемы с конфиденциальностью.

Желательный метод классификации должен быть «невидим» для сети и ее пользователей. Это означает, что она не должна влиять на производительность сети и служб, предоставляемых пользователям. С этой точки зрения, статистическая классификация имеет несомненные преимущества, поскольку не замедляет работу сети, но в тоже время дает относительно точные результаты (решения).

Раздел заканчивается постановкой научной и частных задач исследования.

Во **втором разделе «Контролируемая классификация приложений использующих стек TCP/IP»** решаются частные задачи рационального формирования исходных данных и анализ программного обеспечения, возникающие при разработке системы классификации трафика, используя известные алгоритмы МО. Для получения требуемых исходных данных была использована сеть (Рисунок 1), состоящая из одного компьютера под управлением Windows 10 – локальной рабочей станции, на которой была запущена виртуальная машина с гостевой операционной системой Ubuntu16.04 – виртуальной рабочей станцией, необходимой для функционирования программы tcpdump.

Была создана экспериментальная база данных сетевых приложений WEB (http, https); mail (smtp, imap), Ftp (Ftp-data, Ftp-commands); SSH, Skype, P2P, объединенных в укрупненные группы. Обучающий, общий тестовый и все групповые тестовые дампы являются независимыми друг от друга наборами трафика, что позволяет проводить независимую оценку качества работы классификатора. Предложена архитектура и реализовано специальное

программное обеспечение для формирования исходных данных в задачах классификации различных приложений.

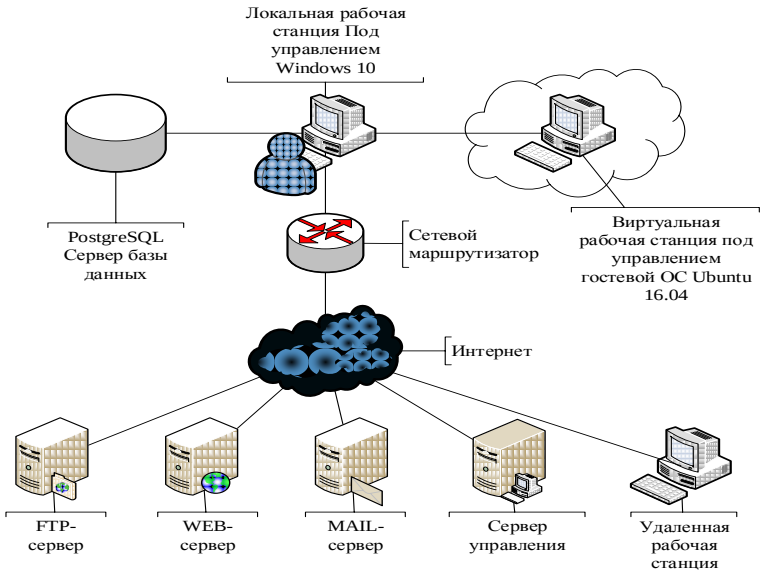


Рисунок 1. Топология исследуемой сети

Полученные экспериментальные результаты измерения с помощью программы tcpdump показали, что распределение наиболее значимых приложений в собранном дампе трафика по размеру в потоках и пакетах (байтах) существенно различаются, что должно учитываться при классификации трафика. Сравнительный анализ существующих снифферов Wireshark, Libtrace, утилит Gt и nDPI, показал достоинства использования утилиты tcpdump в качестве пакетного сниффера, позволяющего эффективно перехватывать данные.

В качестве численных критериев оценки эффективности качества классификаторов в работе использовались такие метрики, как Precision (Точность), Recall (Полнота), F-Measure (F-мера) и AUC (Area Under Curve) – площадь под кривой ROC (Receiver Operating Characteristic- рабочая характеристика приёмника).

Сравнительный анализ методов выбора атрибутов показал, что фильтрационный предпочтителен в случае жестких ограничений вычислительных возможностей. При игнорировании вычислительных затрат предпочтительнее использовать оберточный метод выбора атрибутов, обеспечивающий наивысшую точность классификации; однако последние является «медленным» и не осуществим для использования в приложениях, ограниченных по времени.

Так, если $\boldsymbol{A}$– это признак, $\boldsymbol{C}$– рассматриваемый класс, то энтропия до и после наблюдения признака оценивается выражением (1), при этом изменение

энтропии за счет применения атрибута характеризует информационный выигрыш:

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c), \tag{1}$$

$$H(C|A) = -\sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2 p(c|a).$$

В результате каждому атрибуту $\boldsymbol{a}$ из множества атрибутов $\boldsymbol{A}$ присваивается оценка, основанная на информационном выигрыше между ним самим и классом $\boldsymbol{C}$:

$$IG_i = H(A_i) + H(C) - H(A_i, C). \tag{2}$$

Результатом работы алгоритма InfoGain является ранжирование признаков по их значимости. С помощью него осуществлен выбор числа атрибутов задачи классификации приложений в трафике, позволивший уменьшить общее количество с 21 до 14. Показано, что число атрибутов для классификации не так важно, как выбор рационального алгоритма классификации. Тем не менее, важно избегать атрибутов, которые содержат конкретные значения только для небольшого количества случаев, принадлежащих к конкретному классу, что может привести к лишнему обучению классификатора и, также, не будет возможности идентифицировать неизвестные случаи.

Исследования показали, что для достижения точности классификации не менее 95% достаточно вычислить классификационные атрибуты для образца, содержащего 5…10 последовательных пакетов. При этом тестирование проводилось на наборе данных, различных групп трафика: Skype, FTP, BitTorrent, WEB, SSH. Показано, что точность классификации может быть улучшена до 99,5%, если статистика будет рассчитываться для 35-и последовательных пакетов; дальнейшее увеличение их числа практически не приводит к повышению точности. Поскольку первые несколько пакетов в каждом потоке могут иметь разные характеристики по сравнению с характеристиками остальных пакетов в потоке, целесообразно пропускать первые 10 пакетов в каждом потоке и не включать их в обучающую последовательность.

Сравнение показателей эффективности алгоритмов классификации на этапе обучения производилось при различных размерах обучающих выборок: 30; 60; 120; 180; 300; 600 и 900 тыс. экземпляров соответственно. 66% исходных данных использовались как обучающий набор, остальные 34% – для тестирования и оценки. Показано, что при всех указанных размерах выборок поточная достоверность колеблется в пределах 98…99%, при этом байтовая достоверность практически во всех случаях меньше, чем поточная. При

небольших размерах обучающих выборок (до 60 тыс. экземпляров) достоверность алгоритма *CART* составляет 89%, в то время, как для алгоритма *RF* составляет 91…93% (Рисунок 2).

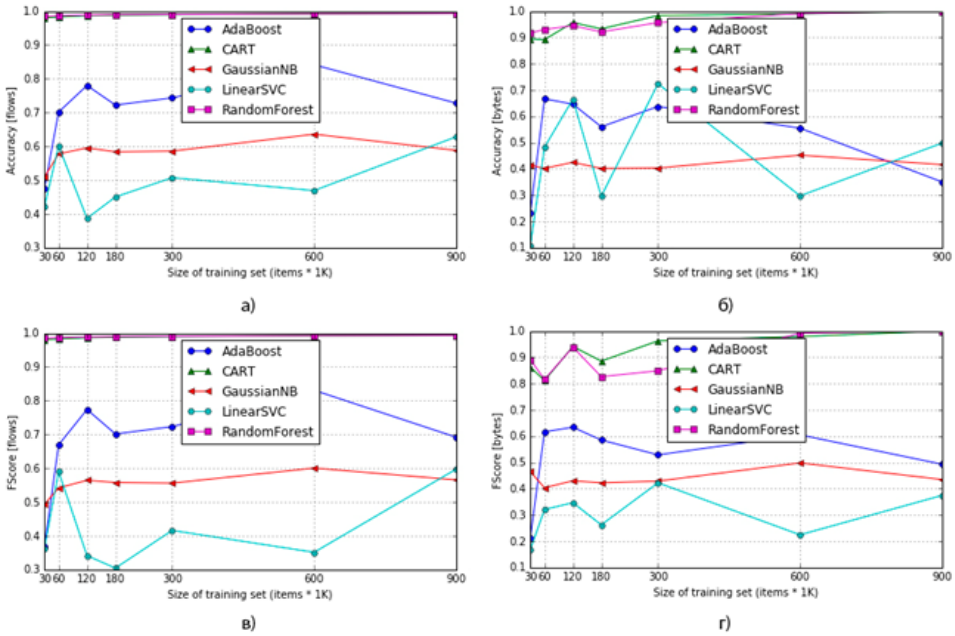


Рисунок 2. Зависимость достоверности и F-меры алгоритмов от размера обучающей выборки: а) поточная достоверность; б) байтовая достоверность; в) поточная F-мера; г) байтовая F-мера

Выявлено, что во всех анализируемых случаях эффективность классификации по потокам выше, чем по байтам.

Результаты классификации на этапе тестирования показали, что алгоритмы МО с использованием «деревьев решений» RF (Рисунок 3) наилучшим образом подходят для классификации при большом количестве классов. Показатели точности классификации с использованием AdaBoost и Bagging указывают на то, что в большинстве случаев объединение множества классификаторов в ансамбль и принятие решения на основе «голосования» позволяет улучшить результаты классификации.

Алгоритмы SVM и Naive Bayes показывают низкую достоверность результатов, что не позволяет, говорить об их применимости для классификации трафика. Хотя алгоритм AdaBoost и использует алгоритм CART для построения базовых классификаторов, его достоверность и F-мера колеблются от 21 до 66%, что также делает его применение для классификации трафика нецелесообразным.

Важной особенностью алгоритма классификации является его способность одинаково хорошо идентифицировать каждый протокол по байтам и по пото-

кам. Из диаграмм, представленных на Рисунке 3 можно видеть, что такой способностью обладает алгоритм CART, показавший наилучшие показатели.

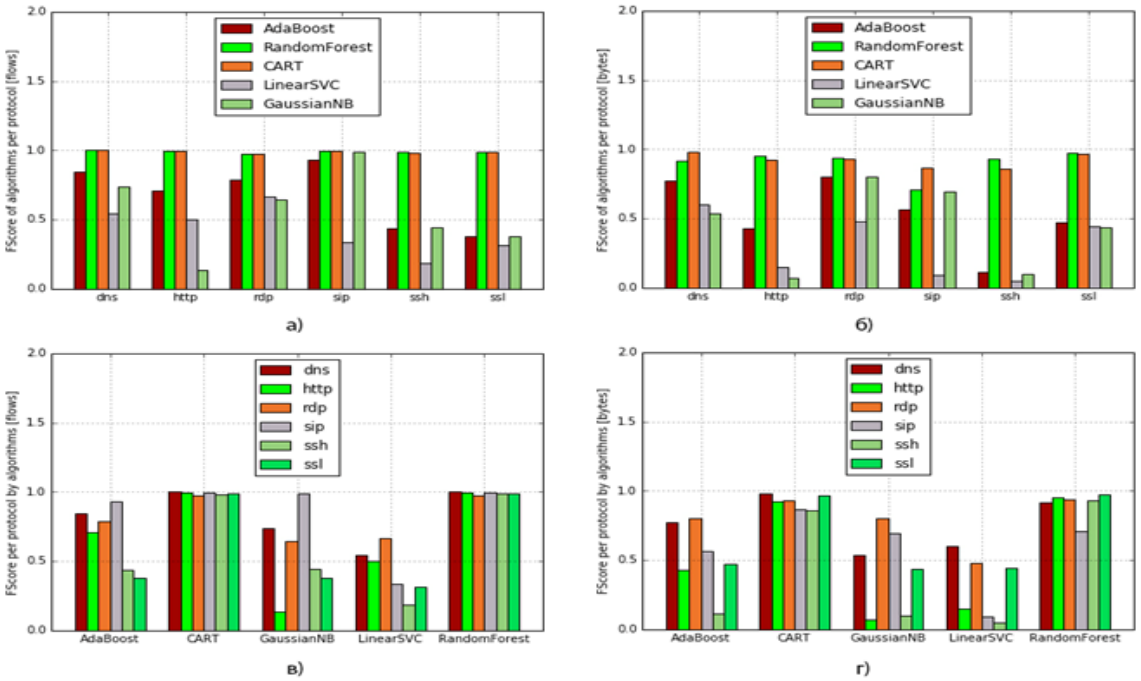


Рисунок 3. Достоверность различных алгоритмов классификации приложений: а) поточная достоверность по протоколам; б) байтовая достоверность по протоколам; в) поточная достоверность протокола по определенному алгоритму; г) байтовая достоверность протокола по определенному алгоритму

Оценка моделей классификации по времени построения показывает, что наихудшим является алгоритм SVM, а наилучшим Naive Bayes. Полученные зависимости достоверности классификации и F-меры показали, что алгоритм CART и Random Forest более предпочтительны, чем алгоритмы SVM, Adaboost и Naive Bayes.

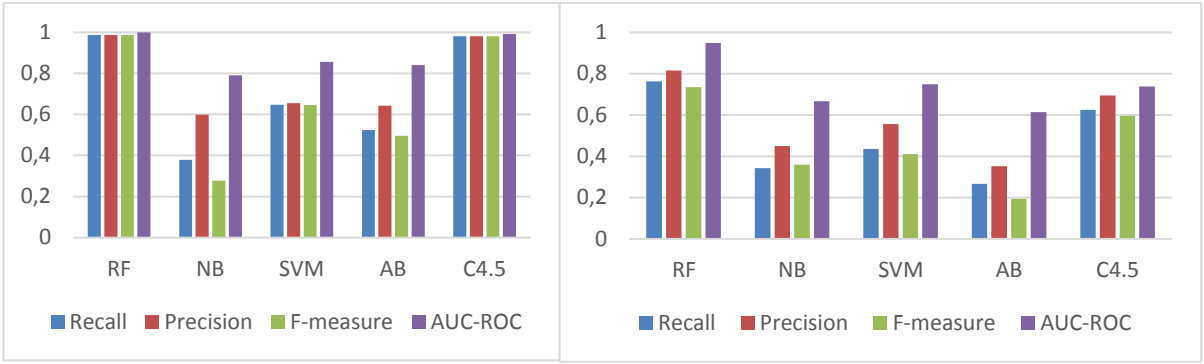
Показано, что во всех анализируемых случаях эффективность классификации по потокам выше, чем по байтам.

В **третьем разделе «Классификация приложений в условиях априорной неопределенности»** исследованы особенности классификации приложений использующих стек TCP/IP в условиях априорной неопределенности. Последняя характеризовалась наличием фонового трафика принадлежащего к классам, которые не участвовали в обучении алгоритма, что значительно ухудшает точность классификации. Влияние неизвестного фонового трафика на качество классификации с использованием МО рассматривалось на примере приложений: HTTP; BitTorrent; Skype; Steam; DNS. Фоновый трафик представлял собой данные, которые не были представлены в

обучающей выборке, однако присутствовали в тестовом наборе: SSL; LLMNR; Quic.

Классификация при наличии фонового трафика показала, что алгоритмы МО с учителем, качество работы которых основывается на полноте и достоверности обучающих выборок данных, не способны определять новые (неизвестные) данные, что приводит к критичным ошибкам. Наличие фонового трафика принадлежащего к классам, которые не участвовали в обучении алгоритма, значительно ухудшает точность классификации.

В качестве примера на рисунке 4 представлены гистограммы усредненных оценок качества классификации для всех рассматриваемых приложений для различных алгоритмов классификации при отсутствии (рис.4 а) и наличии (рис.4 б) фонового трафика.



а) б)

Рисунок 4. Усредненная гистограмма по алгоритмам классификации при: а) отсутствии в наборе данных фона, б) наличии в наборе данных фона

В качестве примера на рисунке 4 представлены гистограммы усредненных оценок качества классификации для всех рассматриваемых приложений для различных алгоритмов классификации при отсутствии (рис.4а) и наличии (рис.4 б) фонового трафика.

Так для приложения Skype снижение величины AUC-ROC для алгоритма *RF* составляет 13,2% , а для C4.5 достигает 35% . Для BitTorrent эти величины составляют: для RF- 5,7% , а для C4.5- 37%. В среднем по всем приложениям снижение составляет для RF 5,2 % а для C4.5 достигает 25%, что обусловлено ложной классификацией фоновых приложений.

Для алгоритма RF величина вероятности ложной классификации FPR при наличии фона возрастает в среднем для всех приложений с величины 0.003 (при отсутствии фона) до величины 0,059, т.е. около 20 раз. Для алгоритма C4.5 увеличение вероятности ложной классификации составляет 18,5 раз. Естественным развитием в данном направлении является применение методов кластеризации, которые предназначены для первичного определения и

разграничения неизвестных типов приложений, а уже затем анализа и классификации.

Однако проведенные исследования алгоритмов кластеризации k-Means и DBSCAN при обучении без учителя, для того же состава трафика протоколов, что и при классификации с учителем показали в целом неудовлетворительные результаты. Рассмотренные алгоритмы неконтролируемого обучения k-Means и DBSCAN значительно уступают в качестве алгоритму обучения с учителем Random Forest. Алгоритм k-Means справляется с кластеризацией потоков сетевого трафика, однако это происходит лишь при условии, что количество кластеров априорно известно. В противном случае качество классификации значительно ухудшается и для приложений HTTP, Skype, DNS, Steam достигает 30%. Алгоритм DBSCAN имеет значительные ошибки в количестве и содержании кластеров ряда анализируемых приложений (классы HTTP , SSL, Steam и Skype оказались разбросаны по многим кластерам).

Повысить эффективность классификации в условиях возможного появления неизвестного фона можно введением дополнительного класса под названием «Неизвестное приложение». Так при введении дополнительного класса величина TPR снизилась для алгоритма RF с 0,987 до 0,964, т.е. не более 2,5%, в то время как величина FPR снизилась на 33%.

На рисунках 5 представлены зависимости средних значений TPR и FPR характеризующие эффективность классификации анализируемых приложений включая введенный в рассмотрение новый класс.

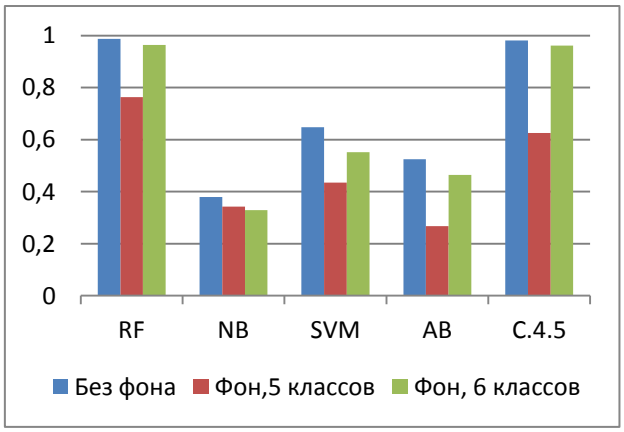


Рисунок 5а. Сравнительная гистограмма параметра True Positive Rate для алгоритмов классификации

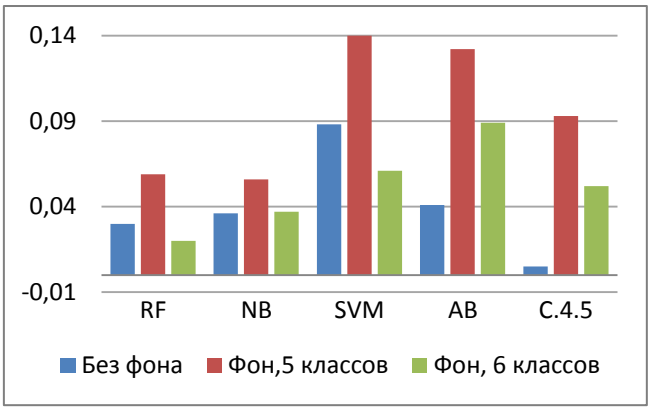
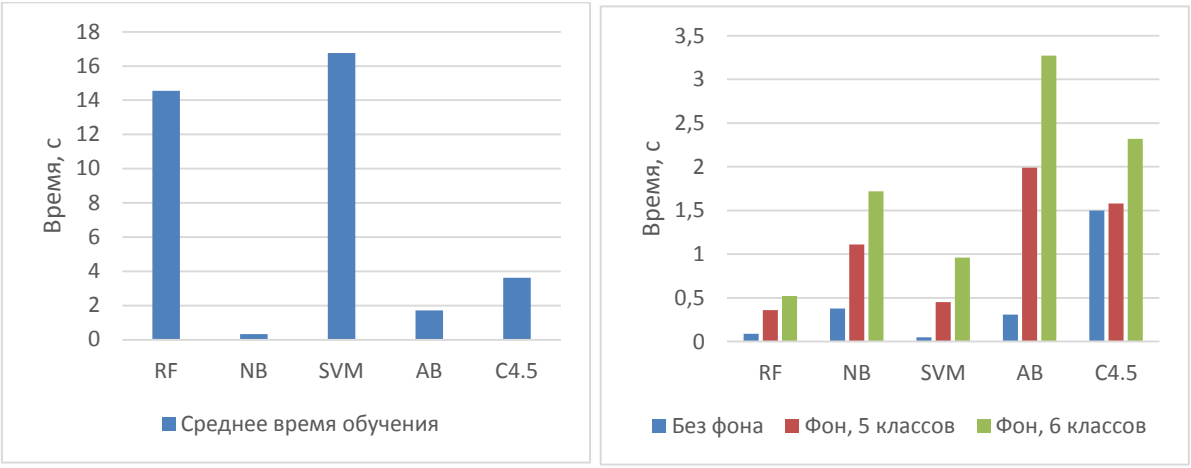


Рисунок 5б. Сравнительная гистограмма параметра False Positive Rate для алгоритмов классификации

Алгоритмы классификации RF и C4.5 показали близкие результаты. Так средняя величина AUC-ROC для всех приложений составляет для алгоритма RF - величину 0,984, а для алгоритма C4.5 – величину 0,985. Однако, как это видно из рисунка 6 , по времени обучения и тестирования они суще-

ственно различаются. Так время тестирования для алгоритма RF в среднем в 4 раза меньше чем у алгоритма C4.5.



а) б)

Рисунок 6. Сравнительная гистограмма времени работы алгоритмов: а) при обучении, б) при тестировании

Важным аспектом классификации приложений в условиях априорной неопределенности является работа алгоритмом в потоковом режиме, при непрерывном поступлении данных измерений. Классификация в условиях нестационарных потоков данных, должна осуществляться в паре с детектором дрейфа («детектором смены концепта»).

С этой целью в работе, при непрерывном поступлении данных измерений предложен алгоритм обнаружения смены концепта наблюдаемого потока базирующийся на оценке статистических характеристиках атрибутов, анализируемых с помощью двух скользящих окон, контролирующих изменение текущих характеристик атрибутов.

Статистический анализ атрибутов анализируемых приложений показал, что наиболее важные из них, связанные с изменением объема трафика имеют экспоненциальный вид. В этих условиях для обнаружения аномальных изменений анализируемых атрибутов предложен алгоритм обнаружения смены концепта наблюдаемого потока $Y = \{y_0, y_1, \dots, y_{N-1}\}$ на базе использования критерия Фишера для выбросов средних значений. Здесь Y_t – значение, элементов потока (приложений и атрибутов) измеренное в момент времени $t \in T = \{0,1, \dots, N-1\}$, а N – размер множества Y. Обнаружение дрейфа анализируемых приложений осуществляется с помощью двух скользящих окон W_1 и W_2, контролирующих изменение текущих статистических характеристик атрибутов. Окно W_1 характеризует статистику атрибутов приложения J в «прошлом»: $Y_1^J = \{y_{ij}^J; j = \overline{1,K}; i = \overline{1;N1}\}$. Окно W_2 характеризует стати-

стику *J*-го приложения «в настоящем»: $Y\ _2^J = \{y_{ij}^J; j = \overline{1,K}; i = \overline{1;N2}\}$. Здесь y_{ij}^J- j-й атрибут, приложения J, для i-го интервала наблюдения; K – количество атрибутов у приложения J; $N_1 -$ длительность окна памяти ; $N_2 -$ длительность окна анализа:и $0 < N_2 < N_1$. Использование «скользящих окон» позволяет решать задачу обнаружения даже незначительных изменений статистических характеристик анализируемых атрибутов в реальном времени. Для обнаружения дрейфа j-го приложения предлагается ввести решающую статистику , на основе которой принимается решение о его наличии, либо отсутствии :

$$R_t^J = \frac{M_{W_2}^J(t)}{M_{W_1}^J(t)} > \lambda, \tag{3}$$

Где $M_{W_1}^J(t) = \frac{1}{N_2 K}\sum_{i=t}^{t-N_2}\sum_{j=1}^{K} y_{ij}^J -$ среднее значение атрибутов J – го приложения в окне W_1;

$$M_{W_2}^J(t) = \frac{1}{N_1 K}\sum_{i=t-N_2}^{t}\sum_{j=1}^{K} y_{ij}^J -$$ среднее значение атрибутов J – го приложения в окне W_2 .

Превышение решающей статистикой R_t^J порогового уровня λ свидетельствует об изменении характеристик наблюдаемых приложений и указывает на необходимость текущего переобучения классификатора. Увеличение порога λ приводит к уменьшению количества ложных срабатываний, однако может приводить либо к пропускам смены концепта, либо к задержкам в обнаружении. Алгоритм обнаружения дрейфа на основе статистики (3), в работе - детектор дрейфа Фишера (ДДФ).

При потоковом режиме классификации анализируется адаптивный алгоритм RF путем введения так называемых «запасных деревьев», обучение которых начинается всякий раз, когда ДДФ зафиксировал наличие смены концепта. Запасные деревья обучаются параллельно со всем ансамблем, но, при этом, не участвуют в предсказании до того момента, пока не будет принято решение о фактическом дрейфе и используемое дерево не перестанет быть эффективным. В этот момент дерево, для которого был зафиксирован дрейф, заменяется ассоциированным с ним запасным деревом.

В качестве примера на рисунке 7 представлены графики значений метрики Recall и Precision полученные в ходе классификации анализируемых приложений для двух алгоритмов обнаружения тренда: ADWIN и ДДФ.

Сравнительный анализ полученных характеристик метрик классификации показывает, что на интервале обучения алгоритм ДДФ обладает большим быстродействием чем ADWIN. В режиме тестирования оба алгоритма показывают близкое качество.

На рисунке 7 представлены графики зависимостей метрики Recall и метрики Precision полученные в ходе классификации анализируемых приложений Instagram и Skype для двух алгоритмов обнаружения тренда: обычно используемого ADWIN и разработанного ДДФ.

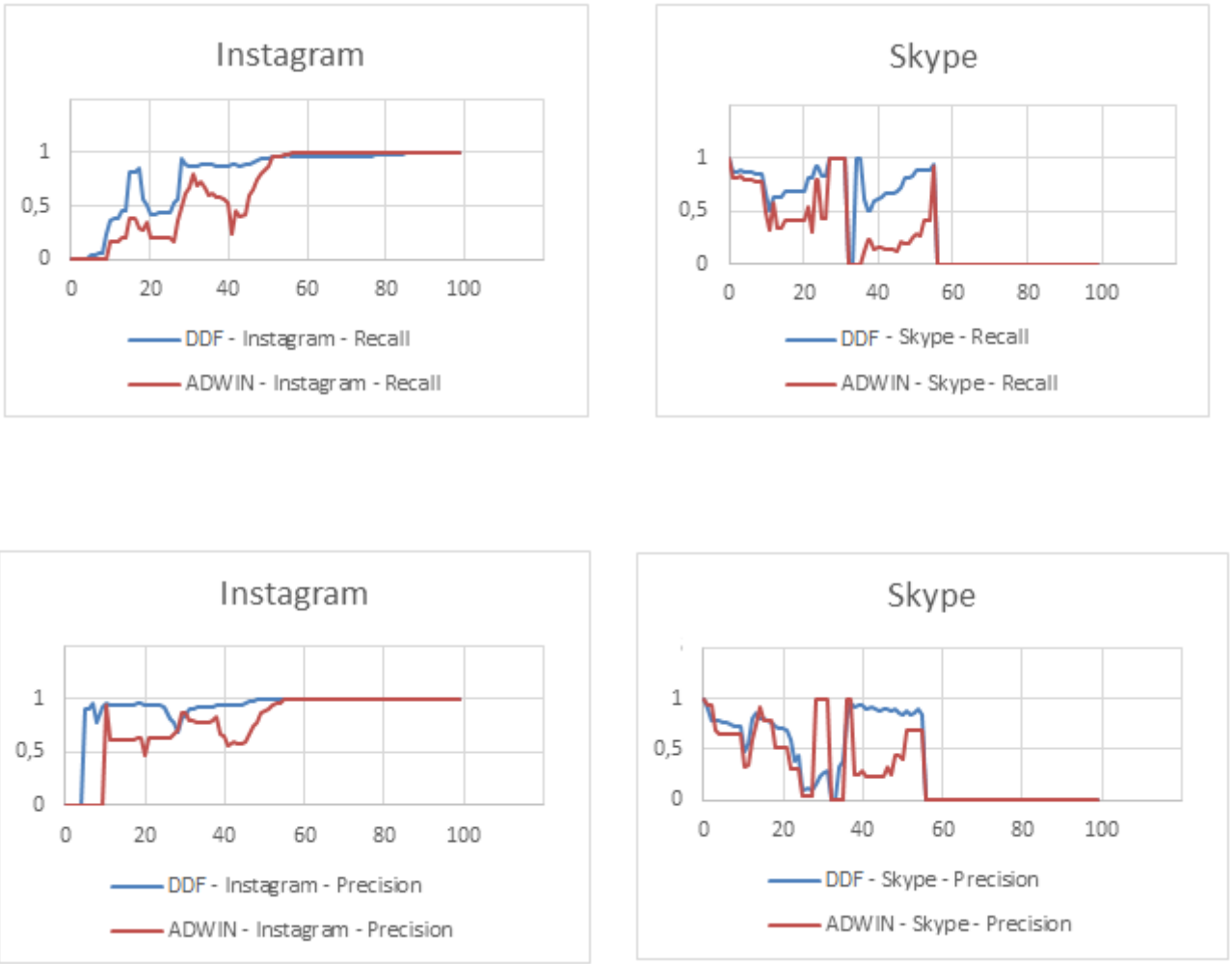


Рисунок 7. Сравнительный анализ качества классификации в потоковом режиме при смещении концепта для приложений Instagram и Skype. (N1=10000; N2=1000; K=6)

Четвёртый раздел «Разработка программного обеспечения для автоматической классификации IP-трафика» посвящен разработке специализированного программного обеспечения, реализующего методы машинного обучения в реальном масштабе времени для классификации IP-трафика. Для этого был выбран объектно-ориентированный язык

программирования Java, основной особенностью которого является то, что все программы при выполнении транслируются в специальный байт-код, выполняемый виртуальной машиной Java. Для разработки программного обеспечения была выбрана технология Java SE8, а для построения графического интерфейса пользователя - JavaFX, содержащая подключаемые библиотеки jNetPcap и Java-ml. В качестве среды разработки была выбрана интегрированная среда разработки IntellijIDEA2016.Подготовка и формирование атрибутов выделялись в отдельную функцию. Последняя не вызывалась каждый раз при добавлении пакета для оптимизации работы программы, поскольку необходимость в формировании значений атрибутов возникала лишь после того как поток признавался закрытым, когда больше пакетов не добавлялось. Разработанный графический интерфейс представляет собой вкладки, на каждой из которых возможно взаимодействие с одним модулем СПО . Для нормальной эксплуатации к разработанному СПО предъявляются следующие требования: операционная система семейства Windows7, Linux, MacOSX и выше; виртуальная машина Java версии JRE8; локально установленная база данных, либо удаленный сервер с установленной базой данных PostgreSQL; библиотека libpcap (поставляются под Windows в составе Winpcap). Разработанное специальное программное обеспечение позволяет:

- выделять потоки и вектора признаков из трафика, независимо от источника, начиная от дампа трафика, либо осуществлять захват с сетевой карточки в реальном времени;
- осуществлять взаимодействие с базой данных (БД) – сохранять выделенный вектор признаков, а также производить SQL-запросы в БД для сохранения потоков в файл набора данных, пригодного для классификации;
- реализовать алгоритмы выделения признаков InfoGain, CFS, Wrapper, а также алгоритмы классификации RF, C4.5, SVM, AdaBoost, Bagging; при этом автоматически проводится оценка введенных показателей.

В СПО реализуется классификация «на лету» за счет созданного «обработчика PCAP» в классе Sniffer и FlowHandler, – то есть проводить автоматическую классификацию в условиях, приближенных к реальным на протяжении всемIP-трафика в системе.

В **заключении** изложены основные результаты проведенных исследований, которые сводятся к следующему:

1. На основе результатов экспериментальных исследований функционирования КС, построенных с использованием телекоммуникационных технологий, создана экспериментальная база данных для задач классификации сетевых приложений: WEB (http, https); mail (smtp, imap), Ftp (Ftp-data, Ftp-

commands); SSH, Skype, P2P. Обоснована рациональная структура и объем обучающей и тестирующей выборок анализируемых сетевых приложений.

2. Показано, что при объеме обучающих выборок 30 ….900 тыс. экземпляров поточная достоверность идентификации сетевых приложений использующих стек протоколов TCP/IP колеблется в пределах 98…99%, при этом байтовая достоверность во всех случаях хуже, чем поточная и не превышает 90% . Предложено сократить число атрибутов анализируемых приложений на 33% , при незначительном (не более чем на 5…7%), снижении достоверности правильной классификации но значительном упрощении процедуры обработки данных.

3. Установлено, что для достижения точности классификации сетевых приложений не менее 95% достаточно вычислить атрибуты для образцов, содержащих 5…10 последовательных пакетов. Результаты тестирования показали, что алгоритмы МО с использованием «деревьев решений» RF и C4.5 наиболее целесообразны для классификации при большом количестве классов.

4. Для преодоления проблемы априорной неопределенности вызванной присутствием фонового трафика в КС , построенных с использованием телекоммуникационных технологий, введен в рассмотрение дополнительный класс «Неизвестное приложение». показавший результаты незначительно уступающие по качественным показателям алгоритмам классификации «с учителем» и являющийся альтернативой методам кластеризации k-Means и DBSCAN. Показано, что для наиболее эффективного алгоритма классификации RF введение дополнительного класса позволило снизить величину ложной классификации (FPR) на 33% , при этом снижение достоверности правильной классификации (TPR) составляет не более 2,5% .

5. Для контроля за изменением текущих статистических характеристик анализируемых атрибутов приложений в КС использующих стек протоколов TCP/IP разработан новый алгоритм обнаружения смещения концепта наблюдаемого потока ДДФ, базирующийся на статистических характеристиках анализируемых атрибутов с помощью двух скользящих окон контролирующих изменение текущих статистических характеристик атрибутов. Показано, что на интервале обучения алгоритм ДДФ обладает более высоким быстродействием чем обычно используемый алгоритм ADWIN.

6. Для обеспечения контроля и диагностики функционирования КС использующих стек протоколов TCP/IP путем автоматической классификации приложений, в том числе в реальном масштабе времени, разработано и реализовано специальное программное обеспечение (СПО), позволяющее выделять потоки и вектора признаков из анализируемых данных независимо от источника, начиная от дампа, а также путем захвата с сетевой карты в реальном времени; осуществлять автоматическую классификацию «на лету».

Основные публикации по теме диссертации

Монография

1. Шелухин О. И.**,** Ерохин С. Д., Ванюшина А. В. Классификация IP-трафика методами машинного обучения / Под ред. профессора О. И. Шелухина. М.: Горячая линия –Телеком, 2018. – 282 с.ISBN 978-5-9912-0719-5

Публикации в научных изданиях, индексируемых в базе Scopus

1. M. G. Gorodnichev ; A. V. Vanushina ; M. S. Moseva ; N. V. Trubnikova Machine Learning in the Tasks of Identifying Unwanted Content // 2019 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF). Saint-Petersburg, Russia. 3-7 June 2019. DOI: 10.1109/ WECONF. 2019.8840130.

Статьи в научных изданиях, входящих в перечень ВАК

1. Ерохин С.Д., Ванюшина А.В Использование алгоритма кластеризации k-Means для идентификации сетевого трафика // Электросвязь. 2018. №12. С.48-49. (0,4 п.л./0,2 п.л.).
2. Ерохин С.Д., Ванюшина А.В. Выбор атрибутов для классификации IP-трафика методами машинного обучения // T-comm: Телекоммуникации и транспорт. 2018. Т.12., №9. С.25-29 (0,9 п.л./0,6 п.л.).
3. Шелухин О.И., Ванюшина А.В., Габисова М.Е. Фильтрация нежелательных приложений Интернет-трафика с использованием алгоритма классификации RandomForest // Вопросы кибербезопасности. 2018. №2(26). С. 44-51. (0,8 п.л./0,4 п.л.).
4. Ерохин С.Д., Ванюшина А.В. Влияние фонового трафика на эффективность классификации приложений методами машинного обучения // T-Comm: Телекоммуникации и транспорт. 2017. Т.11. №12. С. 31-36. (1,1 п.л./0,7 п.л.).
5. Шелухин О.И., Симонян А.Г., Ванюшина А.В. Влияние структуры обучающей выборки на эффективность классификации приложений трафика методами машинного обучения // T-Comm: Телекоммуникации и транспорт. 2017. Т.11. №2. С. 25-31. (0,9п.л./0,5 п.л.).
6. Sheluhin O.I., Simonyan A.G., Vanyushina A.V. Benchmark data formation and software analysis for classification of traffic applications using machine learning methods. T-Comm. 2017. Vol.11. №.1. P. 67-72. (0,9 п.л./0,5 п.л.).
7. Шелухин О.И., Симонян А.Г., Ванюшина А.В. Эффективность алгоритмов выделения атрибутов в задачах классификации приложений при интеллектуальном анализе трафика //Электросвязь. 2016. №11. С.79-85. (1п.л./0,5 п.л.).

Публикации в сборниках трудов международных научно-технических конференций и форумов

8. Ванюшина А.В. Тенденции и проблемы автоматической классификации приложений IP-трафика методами машинного обучения Сборник трудов XII Международной отраслевой научно-технической конференции «Технологии информационного общества». М.: ИД «Медиа Паблишер», 2018. Том.1. С. 371-372. (0,1 п.л.)

9. Ванюшина А.В. Влияние фонового трафика на эффективность классификации приложений методами машинного обучения. //Труды международной научно-технической конференции «Телекоммуникационные и вычислительные системы». М.: Горячая линия-Телеком. 2017. С. 229. (0,1 п.л.).

10. Шелухин О.И., Ванюшина А.В., Калугин Ю.А. Особенности классификации нестационарного потокового трафика методами интеллектуального анализа // Сборник трудов XI Межд. отраслевой научно-технической конференции «Технологии информационного общества». М.: «Медиа Паблишер», 2017. С.268-269. (0,1 п.л./0,04 п.л.)

11. Ванюшина А.В. Классификация приложений в условиях смещения концепта потоков данных // Труды международной научно-технической конференции «Телекоммуникационные и вычислительные системы-2019». - М: Горячая линия – Телеком, 2019. С.216.(0,1 п.л./0,04 п.л.).