На правах рукописи

Усовик Сергей Викторович

ИДЕНТИФИКАЦИЯ ТРАФИКА КОРПОРАТИВНОЙ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ С ПАКЕТНОЙ КОММУТАЦИЕЙ

Специальность 05.13.15
Вычислительные машины, комплексы и компьютерные сети
(технические науки)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2022

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана).

Научный руководитель: **Андреев Арк Михайлович**
кандидат технических наук, доцент кафедры «Компьютерные системы и сети» ФГБОУ ВО «МГТУ им. Н.Э. Баумана»

Официальные оппоненты: **Еременко Владимир Тарасович**
доктор технических наук, профессор, Федеральное государственное бюджетное образовательное учреждение высшего образования «Орловский государственный университет имени И.С. Тургенева», профессор кафедры информационной безопасности

Глухов Антон Викторович
кандидат технических наук, Публичное акционерное общество «Институт электронных управляющих машин им. И.С. Брука», начальник отделения

Ведущая организация: Федеральное государственное казённое военное образовательное учреждение высшего образования «Академия Федеральной службы охраны Российской Федерации»

Защита состоится 30 июня 2022 года в 13 часов 00 минут на заседании диссертационного совета Д 999.216.02 при МАИ и МГТУ им. Н.Э. Баумана по адресу: 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1, зал Ученого совета ГУК МГТУ им. Н.Э. Баумана.

С диссертацией можно ознакомиться в библиотеке МГТУ им. Н.Э. Баумана и на сайте http://www.bmstu.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1, Ученому секретарю диссертационного совета Д 999.216.02

Автореферат разослан «____» _____ 2022 г.

Ученый секретарь диссертационного совета, д.т.н., доцент А.Н. Алфимцев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертации. В настоящее время сферы деятельности, связанные с информационно-телекоммуникационными технологиями, находятся на этапе интенсивного развития. Наряду с ростом числа пользователей появляются новые виды и типы трафика, протоколы информационного взаимодействия. Происходит ускоренное увеличение количества персональных устройств и межмашинных соединений, в том числе, в корпоративном сегменте. Указанные тенденции выдвигают дополнительные требования к проектированию и администрированию сетей передачи корпоративного трафика. Происходит одновременная трансляция множества неоднородных потоков с изменяющейся интенсивностью, отсутствует возможность идентификации и классификации трафика по адресным признакам. Средства шифрования и маскирования трафика используют нестандартные реализации сетевых протоколов. Это отрицательно сказывается на показателях качества функционирования телекоммуникационных сетей. Особенно остро задача обеспечения качества стоит для корпоративных сетей передачи данных. При проектировании таких сетей необходимо иметь представление об объекте управления, под которым понимается динамический процесс передачи трафика. Актуальность диссертационного исследования обусловлена необходимостью разработки механизма точной идентификации разнородного трафика, не имеющего явных признаков классификации.

Проблема идентификации трафика исследовалась в трудах российских и зарубежных ученых, таких как: А. В. Городецкий, О.И. Шелухин, А.М. Тенякшев, А.В. Осин, Ю.Ю. Громов, В.М. Вишневский, Г.А. Урьев, Н.Г. Щербакова, Р.Г. Шыхалиев, Л. Льюнг, T. Lane, T. Karagiannis, A. Dainotti, W. Pescap`e, P.S. Rossi, Sebastian Zander, Thuy Nguyen и другие.

Таким образом, исследования в области идентификации трафика актуальны на настоящий момент и имеют научно-прикладное значение для решения задач в области обеспечения управления, контроля и надежности корпоративных сетей передачи данных.

Объект исследования: трафик корпоративных телекоммуникационных сетей с пакетной коммутацией.

Предмет исследования: алгоритмы и методы идентификации, классификации и кластеризации трафика корпоративных телекоммуникационных сетей с пакетной коммутацией, а также алгоритмы сегментации трафика различных сетевых процессов и источников информации.

Цель работы: разработка и исследование эффективных и быстродействующих алгоритмов идентификации, методов и технологий передачи трафика за счет повышения качества кластеризации и классификации трафика на основе новой математической модели.

Для достижения указанной цели поставлены и решены следующие задачи:

1. Анализ состояния современных корпоративных сетей с пакетной коммутацией на предмет применяемых методов идентификации и моделей трафика, технологий и протоколов передачи информации.

2. Разработка и исследование математической модели трафика корпоративной телекоммуникационной сети с пакетной коммутацией, имеющей уровень новизны и математической строгости.

3. Разработка алгоритма классификации трафика протоколов, используемых в корпоративных телекоммуникационных сетях.

4. Разработка алгоритма скорейшего обнаружения изменения свойств трафика телекоммуникационной сети с пакетной коммутацией, повышающего качество кластеризации трафика.

5. Разработка алгоритма идентификации трафика.

Научная задача: разработка алгоритма идентификации, позволяющего повысить качество сегментации и классификации трафика различных процессов сетевого взаимодействия в условиях высокой степени априорной неопределенности исходных данных.

Методы и математический аппарат исследования: теория вероятностей и математическая статистика, теория информации, математическая теория управления, теория марковских и скрытых марковских случайных процессов.

Научная новизна. В работе получены следующие научные результаты.

1. Предложена новая математическая модель трафика на основе скрытых марковских моделей, отличающаяся от аналогов более высокой точностью аппроксимации и ориентированная на её использование в системах идентификации и классификации трафика в реальном времени.

2. Предложен алгоритм скорейшего обнаружения изменения свойств трафика телекоммуникационной сети с пакетной коммутацией в условиях априорной неопределенности относительно параметров модели, повышающего качество кластеризации трафика.

3. Разработан алгоритм идентификации трафика сети с пакетной коммутацией, позволяющий создать инструмент для решения задачи идентификации трафика в реальном времени в условиях изменения характеристик канала передачи данных.

Теоретическая ценность определяется полученными математическими выражениями, описывающими модель идентификации трафика корпоративной телекоммуникационной сети с пакетной коммутацией, а также алгоритм разделения разнородного трафика.

Практическая значимость состоит в возможности внедрения алгоритма идентификации в системы контроля за изменениями параметров трафика и управления сетевыми процессами. Разработанная модель применима при проектировании вычислительных сетей, при разработке новых устройств, принцип действия которых основан на статистических характеристиках, циркулирующих в сети передачи данных. Предложенные модель и алгоритм могут быть применены при решении задач идентификации вредоносного воздействия для защиты корпоративных телекоммуникационных сетей и передаваемой в них информации.

Внедрение результатов исследований. Результаты работы в виде аналитических и программных средств использованы в ряде разработок в ООО «ТехАргос» и АО «РусБИТех».

Положения, выносимые на защиту:

- математическая модель трафика корпоративной телекоммуникационной сети с пакетной коммутацией;
- алгоритмы классификации трафика протоколов, используемых в корпоративных телекоммуникационных сетях;
- алгоритмы идентификации и сегментации трафика корпоративной телекоммуникационной сети с пакетной коммутацией;

Степень достоверности результатов исследований. Основные результаты, полученные в работе, являются обоснованными на доказательном и экспериментальном уровнях. Достоверность практических результатов достигается за счет большого количества экспериментов при решении задач и использования (наряду с авторским) стандартного программного обеспечения.

Апробация работы. Материалы диссертационной работы докладывались и обсуждались на следующих научных конференциях: «VII Межведомственная конференция «Научно-техническое и информационное обеспечение деятельности спецслужб» (Москва, 2008), «6-я Всероссийская научная конференция «Проблемы развития технологических систем государственной охраны, специальной связи и информации» (Орел, 2009), научных семинарах и заседаниях кафедры компьютерных систем и сетей МГТУ им. Н.Э. Баумана.

Степень соответствия диссертации паспорту научной специальности 05.13.15 «Вычислительные машины, комплексы и компьютерные сети». В рамках диссертационной работы проводились исследования задач идентификации сетевого трафика и сетевых процессов, поиска моментов изменения свойств сетевого потока и выбора решающих правил, исследовалась возможность использования результатов исследований для разработки алгоритмов контроля и диагностики функционирования компьютерных сетей с целью улучшения их технических и эксплуатационных характеристик, что соответствует пунктам 2 и 6 паспорта специальности 05.13.15.

Публикации. По теме диссертации опубликовано шесть работ в ведущих рецензируемых научных изданиях, рекомендованных ВАК РФ. По результатам исследований был получен патент на полезную модель № 94785 «Устройство анализа сетевого трафика».

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы (85 наименований). Полный объем диссертации составляет 176 страниц текста с 37 рисунками и 10 таблицами.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, определены цели и задачи исследования, отражены научная новизна работы и практическая значимость полученных результатов, определяются объект и предмет исследования.

В первой главе описывается состояние современных корпоративных телекоммуникационных сетей с пакетной коммутацией и направления их развития, освещаются актуальные задачи в области их проектирования, эксплуатации и управления. Также определяется место и роль решения задачи идентификации трафика в современной корпоративной телекоммуникационной

сети. Критически оцениваются существующие методы идентификации трафика их возможности и недостатки. Выполняется постановка задачи идентификации трафика при осуществлении управления функционированием корпоративной телекоммуникационной сети с пакетной коммутацией, сформулирована научная задача диссертационного исследования.

В качестве актуальных задач в области проектирования, эксплуатации и управления корпоративными сетями выбраны следующие:

1. Оптимальное распределение нагрузки между равноценными путями на этапе проектирования сети.

2. Прогнозирование изменения параметров трафика (например, интенсивности) во времени.

3. Совместное использование сетевых ресурсов для множества потоков трафика при уменьшении вероятности перегрузок.

Задача идентификации трафика находит свое место в следующих функциональных группах задач управления: обработка ошибок, анализ производительности и надежности, учет работы сети. При отсутствии данных сетевых измерений используется теоретическая модель, полученная при помощи идентификации и математического моделирования характеристик трафика.

При решении задач управления главными параметрами, влияющим на качество, являются своевременность и точность. Своевременность выражается через время реакции - $t_{\text{р}}$. Время $t_{\text{р}}$ складывается из времени идентификации состояния объекта управления - $t_{\text{опр.}}$ и времени выработки непосредственно управляющего воздействия - $t_{\text{упр.}}$, которое приводит объект управления в нужное состояние: $t_{\text{р}} = t_{\text{опр.}} + t_{\text{упр.}}$ Успешное решение задачи идентификации уменьшает время определения состояния объекта управления $t_{\text{опр.}}$ и сокращает время выработки управляющего воздействия $t_{\text{р}}$.

Задача идентификации трафика при осуществлении управления функционированием корпоративной телекоммуникационной сети с пакетной коммутацией формулируется следующим образом: по наблюдаемому трафику $y(t) \in Y$, полученному в результате априорно неизвестного преобразования $A(t)$ абонентского трафика $x(t) \in X$, в общем случае, также неизвестного, построить модель $A^*(t)$, описывающую $y^*(t) \in Y^*$ как максимально приближенный к $y(t) \in Y$ образ неизвестной наблюдателю реализации $x(t) \in X$. В главе представлена структурная схема идентификации телекоммуникационной сети.

Разнообразие методов, решающих общую задачу идентификации, свидетельствует об отсутствии единого подхода к управлению трафиком. Возникает необходимость усовершенствования существующих или разработки новых методов и алгоритмов идентификации трафика сети с пакетной коммутацией, адаптированных для решения задач управления. При этом составной частью процесса идентификации является кластеризация по признакам поведения наблюдаемого трафика. Учитывая это, сформирована цель исследования: повышение качества управления (уменьшение времени реакции и увеличение точности сегментации трафика) при обработке корпоративного трафика телекоммуникационных сетей в условиях отсутствия явных признаков идентификации и классификации. Целевой показатель оценки степени

достижения цели управления трафиком определяется выражением $f(t_p, p_c) \rightarrow max$. Составляющие этого целевого показателя представлены формулами: $t_{\text{п}} = t(Y_n|Y_n) - t(Y_n|Y_m), n, m = 1,2, \dots, k, n \neq m$ и $p_{\text{с}} = \frac{p(Y_n|Y_n)}{p(Y_n|Y_m)}, n, m = 1,2, \dots, k, n \neq m,$ где t_p – время реакции, p_c – вероятность определения принадлежности к заданному классу трафика, Y_i – идентифицируемый класс трафика, k – число классов идентифицируемого трафика, определяемое задачами управления.

Решение целевой задачи управления предполагает выработку управляющего сигнала по пороговым значениям времени h_t и вероятности h_p. Близость h_t и h_p к истинным значениям наступления события определяет соответствие целевому показателю.

Вторая глава посвящена разработке новой математической модели трафика корпоративной телекоммуникационной сети с пакетной коммутацией. В ходе исследования проведен обзор существующих моделей трафика. Для анализа выбраны два наиболее часто используемых классов моделей. Первый класс составляют традиционные модели. Самыми распространенными из них являются модели на основе процессов Пуассона, процессов Бернулли, процессов фазового типа, марковские и скрытые марковские модели. Процессы Пуассона не обладают памятью, что упрощает процесс построения на них очередей обслуживания трафиковых поступлений.

В моделях на основе процессов Бернулли вероятность поступлений в любом временном слоте p не зависит от любого другого поступления. Модели на основе процессов Бернулли обладают аналогичными преимуществами, что и пуассоновские модели.

В моделях на основе фазового типа интервалы между поступлениями пакетов описываются процессом поглощения в непрерывном марковском процессе. Основным преимуществом использования моделей на основе процессов фазового типа является то, что любое распределение $\{\tau_n\}$ может быть сколь угодно близко аппроксимировано процессами фазового типа.

Перечисленные модели не учитывают пульсирующий характер трафика. Корреляционная функция процесса $\{\tau_n\}$ обращается в нуль одинаково для всех ненулевых задержек. В силу указанных причин возобновляющиеся модели не применяются для моделирования трафика в широкополосных сетях, поскольку это приведет к значительному ухудшению характеристик сети.

Марковские и скрытые марковские модели основаны на свойстве, которое говорит о том, что будущее состояние зависит от текущего и не зависит от предыдущих состояний или времени, проведенного в текущем состоянии. Марковская модель трафика описывается пространством состояний $S = \{s_1, s_2, \dots, s_M\}$, а также матрицей переходных вероятностей $\|p_{ij}\|$ в дискретном случае или временем пребывания в определенном состоянии в непрерывном случае. При использовании скрытых марковских моделей наблюдения порождены сменой скрытых состояний. Существенным ограничением использования марковских моделей является необходимость априорно описывать структуру процесса, пространство состояний, число состояний и механизм перехода между состояниями.

Второй класс составляют фрактальные модели трафика. Из них можно выделить модели на основе распределения Парето, фрактального броуновского движения, фрактального гауссовского шума, фрактальных точечных процессов, на основе ARIMA-процессов.

Модели на основе распределения Парето описываются распределением вида: $\omega(x) = \frac{\alpha k^{\alpha}}{(x)^{\alpha+1}}$; $\alpha > 0$; $k > 0$; $x > 0$, где α – параметр формы, k – нижний граничный параметр, т.е. минимальное значение для случайной переменной x. Генерация осуществляется при помощи обратной функции от интегральной функции распределения $y = F^{-1} = \frac{k}{\sqrt[\alpha]{1-rnd}}$, где rnd – случайная переменная, равномерно распределенная на интервале (0;1).

Модели на основе фрактального броуновского движения описываются случайным процессом, начинающимся в нуле координат, приращения которого на непересекающихся интервалах времени независимы и имеют гауссовское распределение. Плотность распределения вероятностей координаты частицы $X(t = n\tau) = \sum_{i=1}^{n} \xi_i$ имеет вид: $\omega(\Delta X) = \frac{1}{(2\pi k_D|\Delta t|)^{0,5}} exp\left\{-\frac{(\Delta X)^2}{2k_D|\Delta t|}\right\}$, где $\Delta X = X(t) - X(t_0)$; $\Delta t = t - t_0$; k_D – коэффициент диффузии.

Модели на основе фрактального гауссовского шума описываются стационарным процессом приращений X_H фрактального броуновского движения: $X_H = \{X_H(t) = B_H(t+1) - B_H(t): t \geq 0\}$.

Модели на основе фрактальных точечных процессов являются асимптотически самоподобными процессами второго порядка, которые характеризуются спектральной плотностью мощности: $S(\omega) = \int_{-\infty}^{\infty} R(\tau)e^{-j\omega\tau}d\tau$; индексом дисперсии отчетов: $(T) = (\lambda T)^{-1}\int_{-T}^{T}(T - |\tau|)[R(\tau) - \lambda^2]d\tau$; корреляционной функцией: $R(k;T) = \int_{-T}^{T}(T - |\tau|)[R(kT + \tau) - \lambda^2]d\tau$.

Модели на основе ARIMA-процессов описывают класс нестационарных рядов $\{Y(t): t \in Z\}$, которые проявляют однородность в отличие от их локального уровня или тренда. Такие ряды описываются обобщенным авторегрессионным оператором $\phi(B)$: $(B) = \alpha(B)(1 - B)^d$, где $\alpha(B) = 1 - \alpha(1)B^1 - \alpha(2)B^2 - \cdots - \alpha(p)B^p$ – многочлен p-степени, d – кратность корня при $B = 1$.

Общим недостатком применения фрактальных моделей является сложность используемого математического аппарата, вследствие чего в современных работах по исследованию трафика продолжают применять традиционные модели трафика. Это делает привлекательным для исследований использование класса моделей, основанных на марковских случайных процессах (MMPP-модель, модель на основе dMMPP-процессов и др.).

Математическая постановка задачи на построение модели трафика сводится к описанию класса трафика при помощи количественных $N(t)$ и временных τ_n параметров, определяющих функциональную зависимость $q(\omega)$: $\Omega \to Y$, где функция $q(\omega)$: $\Omega \to Y$ ставит в соответствие каждому объекту $\omega \in \Omega$ его образ $q(\omega) \in Y$, непосредственно воспринимаемый наблюдателем. Пространство наблюдений Y определяется последовательностями однородных наблюдений. В соответствии с математической постановкой задачи на построение математической модели и произведенным анализом существующих

моделей телекоммуникационных сетей разработана модель телекоммуникационной сети с пакетной коммутацией.

Эта модель представляет собой набор параметров $\{A, \theta_i, i\}$, где A – матрица переходных вероятностей скрытой марковской цепи, управляющей процессом поступления трафика; $\theta_i = (p_i, \lambda_i)$ – вектор параметров смеси пуассоновских распределений, где p_i – вероятность компоненты смеси, λ_i – интенсивность распределения Пуассона (компоненты смеси); i – число компонент смеси. Аппроксимация эмпирической плотности распределения вероятностей смесью распределений Пуассона приведен на рисунке 1.

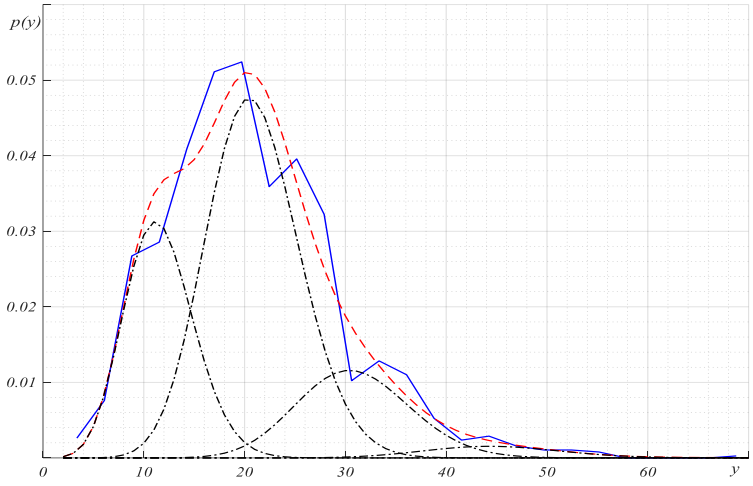

Рисунок 1. Аппроксимация эмпирической плотности распределения вероятностей смесью распределений Пуассона

Отличие данной модели трафика от предложенных ранее заключается в том, что введено число компонент смеси i в виде отдельного параметра благодаря возможности получения его аналитически, при помощи использования алгоритма автоматического определения числа компонент ARD EM. Идея алгоритма состоит в использовании на начальном этапе заведомо избыточного количества компонент смеси с дальнейшим определением релевантных компонент с помощью максимизации обоснованности. Устанавливается начальное число компонент смеси $K = \sqrt{N}$, где N – число наблюдений. Априорное распределение весов смеси определяется выражением:

$$p(\omega_i|\alpha_i) = \sqrt{\frac{\alpha_i}{2\pi}} exp\left(-\frac{1}{2}\omega_i^2\alpha_i\right).$$

Алгоритм ARD EM предполагает оптимизацию функционала: $\Xi_{ML} = (\Theta_{ML}, \omega_{ML}) = argmax_{(\Theta|\omega)} P(Y|\Theta, \omega)p(\omega|\alpha)$, который описывает метод максимального правдоподобия. В данном выражении параметр $\Theta = \{\theta_j\}_{j=1}^K$ описывает совокупность компонент смеси, параметр $\omega = \{\omega_j\}_{j=1}^K$ описывает веса соответствующих компонент. Совокупность указанных параметров обозначается $\Xi = \{\Theta, \omega\} = \{\theta_j, \omega_j\}_{j=1}^K$. Параметр α_i подбирается с помощью максимизации обоснованности:

$$(Y|\Theta_{ML}(\alpha), \alpha) = \int p(Y|\Theta_{ML}(\alpha), \omega)p(\omega|\alpha)d\omega \to \max_{\alpha}.$$

При отнесении наблюдения к определенному состоянию скрытой марковской цепи вывод делается на основании большего значения вероятности одной из компонент смеси $s = s_l \mid p(y_j|\theta_l) > p(y_j|\theta_k), \forall\, \theta_k = \{\theta_1, \theta_2, \dots, \theta_n\}, k \neq l$, однако, чем ближе значения $p(y_j|\theta_l)$ и $p(y_j|\theta_k)$, тем большая вероятность (риск) принять неправильное решение относительно состояния марковской модели. Следовательно, одним из критериев предпочтения выбора модели является

уменьшение риска принятия ошибочного решения относительно последовательности состояний марковской модели, описывающей наблюдения, что приведет к неверному построению марковской цепи A и, как результат, ошибочному построению функции $q(\omega)\colon \Omega \to Y$. Для определения риска принятия неверного решения в случаях с моделями, описываемыми смесью нормальных и пуассоновских распределений, вводится критерий близости значений вероятностей $p(y_j|\theta_l)$ и $p(y_j|\theta_k)$ при определении состояния s_l из i возможных состояний модели: $R = \min\limits_{1\le l,k\le i}(p(y_j|\theta_l) - p(y_j|\theta_k))$, $l \neq k$. Полученные в ходе построения моделей значения R группируются по частоте появления N_R. При использовании смеси плотностей пуассоновских распределений при верном определении числа компонент смеси риск принятия неправильного решения относительно принадлежности наблюдения определенному состоянию скрытой марковской модели ниже. Следовательно, предпочтительным является описание модели трафика в виде смеси распределений Пуассона и построенной на основе его наблюдения матрицы переходных вероятностей скрытой марковской цепи A.

Третья глава посвящена анализу и исследованию свойств предложенной математической модели трафика корпоративной сети с пакетной коммутацией. Модель описывает суммарный трафик в канале передачи данных. Суммарный трафик подчиняется тем же принципам передачи, что и трафик отдельного абонента или приложения. В главе исследована модель $\{A, \theta_i, i\}$ и разработан алгоритм идентификации пользовательского трафика приложений.

В работе исследованы протоколы, наиболее распространенные в корпоративных сетях: HTTP, SMTP, FTP. Полученные результаты показывают, что модели трафика протоколов $\{A, \theta_i, i\}$, где $\theta_i = (p_i, \lambda_i)$, рассчитаны с применением алгоритма ARD EM для получения числа компонент смеси распределений i, имеют максимальное значение функции правдоподобия. Таким образом, трафик рассмотренных приложений описывается моделью $\{A, \theta_i, i\}$, где $\theta_i = (p_i, \lambda_i)$ с автоматическим определением числа компонент i по алгоритму ARD EM. Представленный в главе алгоритм позволяет автоматически классифицировать трафик протоколов. Качество классификации оценено по метрикам точности и полноты. В Таблице 1 представлены результаты оценки качества классификации.

Таблица 1

Результаты классификации трафика протоколов по значению i

Наименование протокола	Метрики оценки качества классификации	
	Точность ($Precision$)	Полнота ($Recall$)
HTTP	0,8	0,6
FTP	0,66	0,4
SMTP	0,7	0,5
NNTP	0,66	0,5

Ошибки классификации трафика сетевых протоколов связаны с тем, что в результате применения алгоритма ARD EM для трафика разных протоколов возможно получение одинаковых значений i.

Дополнительно исследованы параметры A и θ_i модели $\{A, \theta_i, i\}$ на возможность использования их для классификации трафика протоколов. Результаты исследований показывают, что трафик протоколов различим как на уровне средних значений $\{\lambda_i\}$ и $\{p_i\}$ в каждой из компонент смеси, так и на уровне разброса значений и частоты их появления.

Следовательно, для классификации трафика применим метод разделения наблюдений по пороговому значению.

Принадлежность $\theta_i = (p_i, \lambda_i)$ определенному состоянию i, а также переходы между состояниями описываются матрицей переходных вероятностей A. Отнесение трафика к определенному приложению по наблюдению за θ_i определяется различным поведением A. Вероятность p_i появления значения λ_i, принадлежащего определенной компоненте смеси, зависит от ненулевых значений переходных вероятностей p_{ij} матрицы $A = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix}$. Для оценки этой характеристики использована вероятность того, что матрица не имеет элементов, равных нулю. Значения этой вероятности вычисляются по формуле: $P_{p_{ij}\neq 0} = \frac{N_{p_{ij}\neq 0}}{N_{p_{ij}}}$. Параметры A и $\theta_i = (p_i, \lambda_i)$ дополняют параметр числа компонент модели i для классификации трафика приложений. В главе представлен алгоритм классификации трафика (рисунок 2). Результаты оценки качества классификации рассмотренных реализаций на каждом этапе представлены в Таблице 2.

Таблица 2

Оценка качества последовательной классификации трафика

Наименование протокола	Метрики оценки качества классификации трафика приложений					
	Точность ($Precision$)			Полнота ($Recall$)		
	i	$\theta_i = (p_i, \lambda_i)$	A	i	$\theta_i = (p_i, \lambda_i)$	A
HTTP	0,8	0,9	0,94	0,6	0,7	0,73
FTP	0,66	0,8	0,83	0,4	0,5	0,52
SMTP	0,7	0,73	0,8	0,5	0,6	0,7
NNTP	0,66	0,7	0,74	0,5	0,65	0,8

Таким образом, полнота и точность классификации повышаются при использовании значений параметров A и θ_i модели. Исследована зависимость параметров модели $\{A, \theta_i, i\}$ от загруженности канала передачи данных. Для этого произведена оценка корреляции между значениями интервалов времени T между соседними пакетами в реализации трафика и значениями размеров пакетов данных Y других пользователей и сетевых устройств за время T по следующей формуле:

$$r = \frac{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(T_i - \overline{T})^2 \sum_{j=1}^{n}(Y_j - \overline{Y})^2}}.$$

Результаты показали отсутствие корреляции между значениями межпакетных интервалов и числом передаваемых пакетов данных других пользователей и сетевых устройств.

Оценено влияние загруженности канала передачи данных на параметры модели трафика. Результаты исследований свидетельствуют об отсутствии указанного влияния в наблюдаемых реализациях трафика сетевых протоколов.

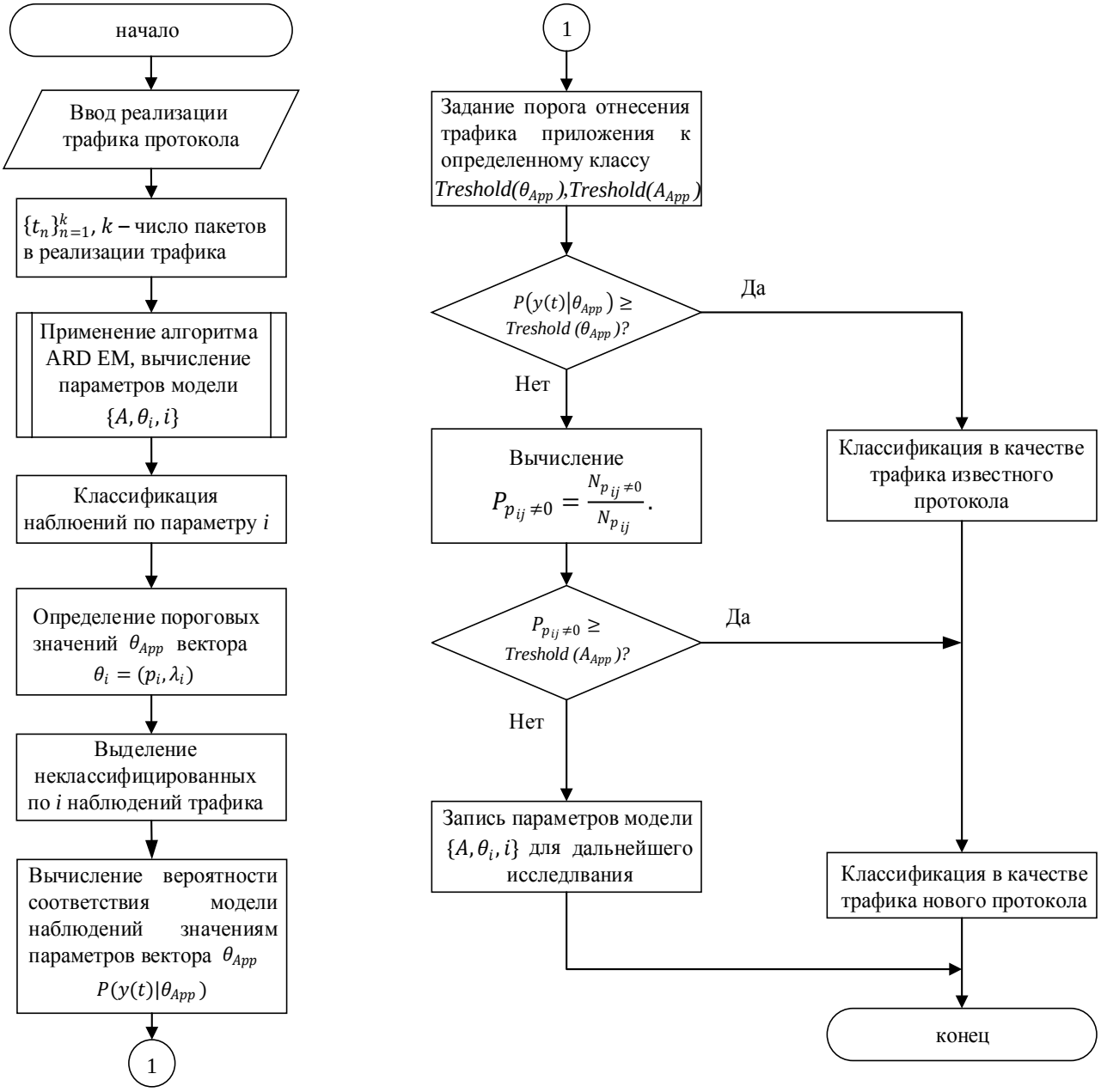

Рисунок 2. Алгоритм классификации трафика приложений на основе значений параметров A и $\theta_i = (p_i, \lambda_i)$

Следовательно, полученные параметры модели конкретного сетевого приложения и их взаимные связи возможно использовать в качестве идентификационных признаков. Значения параметров модели $\{A, \theta_i, i\}$ характеризуют индивидуальные особенности трафика приложений и работы пользователей.

Четвертая глава посвящена разработке алгоритма идентификации трафика корпоративной сети с пакетной коммутацией. Проведен структурный анализ трафика и описано его представление в виде статистически мультиплексированного потока данных. Групповой трафик, полученный способом статистического мультиплексирования, обладает дополнительным уровнем априорной неопределенности относительно длительностей промежутков времени (timeslot), выделяемых каждому процессу. Сетевой

процесс обладает определенными значениями вектора параметров модели $\psi_j = \{A, \theta_i, i\}$. Множество ψ_j образует множество возможных состояний группового сигнала $\Psi = \{\psi_1, \psi_2, \dots, \psi_n\}$, где n – общее число сетевых процессов. В системе статистического уплотнения по времени обозначенная априорная неопределенность выражается следующим образом: $p_{\psi_k|\psi_l}(t) \neq 0$, где ψ_k и ψ_l – состояния сетевых процессов, сменяющих друг друга в реализации группового сигнала. Сформулирована задача обнаружения изменения свойств трафика как случайного процесса. Пусть $t \in (t_0, t_N)$, где (t_0, t_N) – фрагмент временной оси и на этом фрагменте определен случайный процесс $X(t, \Psi)$, в котором элементы множества $\Psi = \{\psi_1, \psi_2, \dots, \psi_n\}$ скачкообразно сменяют друг друга. При этом $x(t, \psi_i)$ – наблюдаемая реализация процесса $X(t, \Psi)$, где $X(t, \Psi) = \{x(t, \psi_i)\}$, $i = 1, \dots, n$. Количество изменений K в общем случае неизвестно. Наблюдаемые на интервалах времени (t_0, t_1), (t_1, t_2), … , (t_{K-1}, t_K), (t_K, t_N) реализации случайного процесса $x(\psi_i)|_{t_0}^{t_1}$, $x(\psi_j)|_{t_1}^{t_2}, \dots$, $x(\psi_k)|_{t_K}^{t_{K-1}}$, $x(\psi_l)|_{t_K}^{t_N}$ удовлетворяют условию $\psi_m|_{t_{k-1}}^{t_k} \neq \psi_n|_{t_k}^{t_n}$. Требуется по реализации $x(t, \psi_i)$ случайного процесса $X(t, \Psi)$ на интервале времени (t_0, t_N), определить число K моментов скачкообразного изменения параметров процесса $\Psi = \{\psi_1, \psi_2, \dots, \psi_n\}$ и оценить моменты $t_1, t_2, \dots, t_K$.

Для определения изменения свойств наблюдаемого трафика применен алгоритм последовательного обнаружения изменения свойств (разладки) трафика. В этом случае скорейшим образом определяется момент изменения свойств случайного процесса при заданной вероятности ошибок первого рода α (ложная тревога) при помощи алгоритма куммулятивных сумм. Процедура выбора гипотез согласно алгоритма АКС записывается в виде: $\max\limits_{1 \le \tau \ge N} S_\tau^N \left(\theta^0, \hat{v}_N(\tau)\right) \underset{H_1}{\overset{H_0}{\lessgtr}} h$. Гипотезы H_0 и H_1 соответствуют различным значениям параметра случайного процесса, изменение которого контролируется. Порог h зависит от разности $v = \theta^1 - \theta^0$. Для алгоритма поиска разладки минимальный объем данных наблюдений определяется путем многократного применения процедуры получения оценки $\hat{\theta}^0$ при увеличении числа наблюдений O. Применением итеративных процедур вычисляется максимально правдоподобная оценка $\hat{\theta}^{L_{max}} = \left(\hat{p}_k^{L_{max}}, \hat{\lambda}_k^{L_{max}}\right)$, $k = Q$ каждой реализации. Процедура повторяется для большого количества n реализаций O. Оценки параметров $\hat{p}$ и $\hat{\lambda}$, последовательно получаемых векторов $\hat{\theta}_1^{L_{max}}$, $\hat{\theta}_2^{L_{max}}, \dots$, $\hat{\theta}_n^{L_{max}}$, стремятся к математическим ожиданиям μ^p, μ^λ, полученным для числа наблюдений n. Соответственно, добавление новых наблюдений o для получения $\hat{\theta}^0$ необходимо остановить, когда параметры $\hat{p}_k$, $\hat{\lambda}_k$, $k = Q$ начинают соответствовать гауссовским распределениям. Количество N этих наблюдений $o_1 o_2 \dots o_N$ определяет минимальный объем наблюдений для применения алгоритма поиска разладки. В качестве значений, к которым стремятся значения соответствующих компонентов θ, выступают значения математических ожиданий μ_k^p, μ_k^λ, полученные на тестовых реализациях. В главе представлены изменения значений параметров вектора θ и определены границы h_λ и h_p, после

достижения которых выполняется равенство оценок параметров $\mu_k^{\hat{\lambda}}$ и $\mu_k^{\hat{p}}$ и истинных значений μ_k^{λ} и μ_k^{p} соответственно.

Модель трафика в групповом канале $\Psi_{\text{гр}}$ описывается следующим образом:

$$\Psi_{\text{гр}} = \begin{cases} \Psi^0 = \{A^0, \theta^0, Q^0\}, & A^0 = \{a_{ij}^0\}, 1 \le i,j \le Q^0, \quad t \le \tau, \\ \Psi^1 = \{A^1, \theta^1, Q^1\}, & A^1 = \{a_{ij}^1\}, 1 \le i,j \le Q^1, \quad t > \tau. \end{cases}$$

Ψ^0 и Ψ^1 последовательно сменяющие друг друга модели сетевых процессов, при этом параметры Ψ^0 определены и $\theta^0 \neq \theta^1, A^0 \neq A^1$. В момент времени $t = \tau$ трафик, описываемый моделью Ψ^0, сменяется трафиком, описываемым моделью $\Psi^1 = \{A^1, \theta^1, Q^1\}$ с априорно неизвестными параметрами. Трафик Ψ^0 и Ψ^1 порождает последовательность наблюдений O. Логарифм отношения правдоподобия для проверки гипотез о наличии H_0 и отсутствии H_1 разладки представлен следующей формулой:

$$\Lambda(k) = \frac{\nu_O}{\sigma_O^2} \sum_{m=k}^{M} \left(O_m - \mu^0 - \frac{\nu_O}{2} \right),$$

где O_m – текущие наблюдения; величина $\nu_O = \mu^0 - \mu^1$ с учетом знака есть величина скачка функции, представленной составляющими $O_1, O_2, \dots, O_M$ в анализируемой точке. В зависимости от параметра момент изменения свойств которого необходимо определить под O_m понимается значение $\hat{\lambda}_m$ или $\hat{p}_m$. Для логарифма отношения правдоподобия оценка максимального правдоподобия номера точки $k_{\text{оц.}}$, в которой выполняется гипотеза о соответствии наблюдений модели Ψ^1, записывается в следующем виде:

$$k_{\text{оц.}} = arg \max_k \nu_O \sum_{m=k}^{M} \left(O_m - \mu^0 - \frac{\nu_O}{2} \right).$$

Правило принятия решения о наличии скачка:

$$r: g_k = \Lambda\left(k_{\text{оц.}}\right) = \max_k \nu_O \sum_{m=k}^{M} \left(O_m - \mu^0 - -\frac{\nu_O}{2} \right) \underset{\Psi^1}{\overset{\Psi^0}{\lessgtr}} h,$$

где $t_1 < \tau$, h - порог, выбираемый с учетом рисков принятия ошибочного решения. В условиях априорной неопределенности, когда величина обнаруживаемого скачка ν_O не задана, она является неизвестным параметром функции правдоподобия $\Lambda(k, \nu_O)$. Фиксируя k, получается оценка максимального правдоподобия величины ν_O:

$$\nu_{O\,\text{оц.}} = arg \max_{\nu_0} \nu_O \sum_{m=k}^{M} \left(O_m - \mu^0 - -\frac{\nu_O}{2} \right) = \frac{1}{M-k+1} \sum_{m=k}^{M} (O_m - \mu^0).$$

Оценка максимального правдоподобия номера точки $k_{\text{оц.}}$, в которой выполняется гипотеза о соответствии наблюдений модели Ψ^1, записывается в следующем виде:

$$k_{\text{оц.}} = arg \max_k \frac{1}{2(M-k+1)} \left[\sum_{m=k}^{M} (O_m - \mu^0) \right]^2.$$

Правило принятия решения при оценке координаты $k_{\text{оц.}}$ имеет вид:

$$r: g_k = \Lambda\left(k_{\text{оц.}}, \nu_{O_{\text{оц.}}}\right) == \max_k \frac{1}{2(M-k+1)}\left[\sum_{m=k}^{M}(O_m - \mu^0)\right]^2 \underset{\Psi^1}{\overset{\Psi^0}{\lessgtr}} h.$$

Для моделей Ψ^0 и Ψ^1 необходимо контролировать изменение двух параметров: A и λ. Изменение A влияет на изменение набора параметров p_k вектора θ. Существует три варианта изменения параметров Ψ при $t > \tau$: 1) $A^0 \neq A^1, \{\lambda_i^0\} = \{\lambda_i^1\}$; 2) $A^0 = A^1, \{\lambda_i^0\} \neq \{\lambda_i^1\}$; 3) $A^0 \neq A^1, \{\lambda_i^0\} \neq \{\lambda_i^1\}$. Для определения количественной оценки вероятности появления конкретной последовательности наблюдений $o_1\ o_2\ \dots\ o_t$ до момента t и состояния в этот момент для определенной модели Ψ^i вводится прямая переменная $\alpha_t(i)$, определяемая выражением $\alpha_t(i) = P(o_1 o_2 \dots o_t, q_t = S_i|\Psi^i)$. Далее выполняется алгоритм прямого-обратного хода. Для скорейшего обнаружения момента разладки τ применяется АКС, где в качестве входного параметра АКС используется $\mu\left(P\left(O_{t_1}|\Psi^0\right)\right)$, $t_1 \leq \tau$. Эта величина получена путем, оценки параметров модели Ψ^0. По мере поступления наблюдений O применяется алгоритм прямого-обратного хода для получения значений $P(O_t|\Psi)$.

Оценка максимального правдоподобия номера точки $k_{\text{оц.}}$, в которой выполняется гипотеза о соответствии наблюдений модели Ψ^1, записывается в следующем виде: $k_{\text{оц.}} = arg\max_k \frac{1}{2(M-k+1)}\left[\sum_{m=k}^{M}\left(O_m - \mu\left(P\left(O_{t_1}|\Psi^0\right)\right)\right)\right]^2, t_1 < \tau$. Правило принятия решения r при оценке координаты $k_{\text{оц.}}$ имеет вид:

$$r: g_k = \max_k \frac{1}{2(M-k+1)}\left[\sum_{m=k}^{M}\left(O_m - \mu\left(P\left(O_{t_1}|\Psi^i\right)\right)\right)\right]^2 \underset{\Psi^1}{\overset{\Psi^0}{\lessgtr}} h,$$

где $t_1 < \tau$, h - порог, выбираемый с учетом риска принятия ошибочного решения. Максимум оценки максимального правдоподобия номера точки $k_{\text{оц}}$ определяется при анализе $P_k(O|\Psi)$, $k = 1,2,\dots,n$. Наибольшее значение g_k, по сравнению с предысторией, наблюдается при рассмотрении $P_k(O|\Psi)$, $k = 1,2,\dots,n$ при $\tau_{\text{оц.}}$, наиболее близком к $\tau_{\text{ист}}$ (рисунок 3). В главе построен итоговый алгоритм идентификации трафика (рисунок 4).

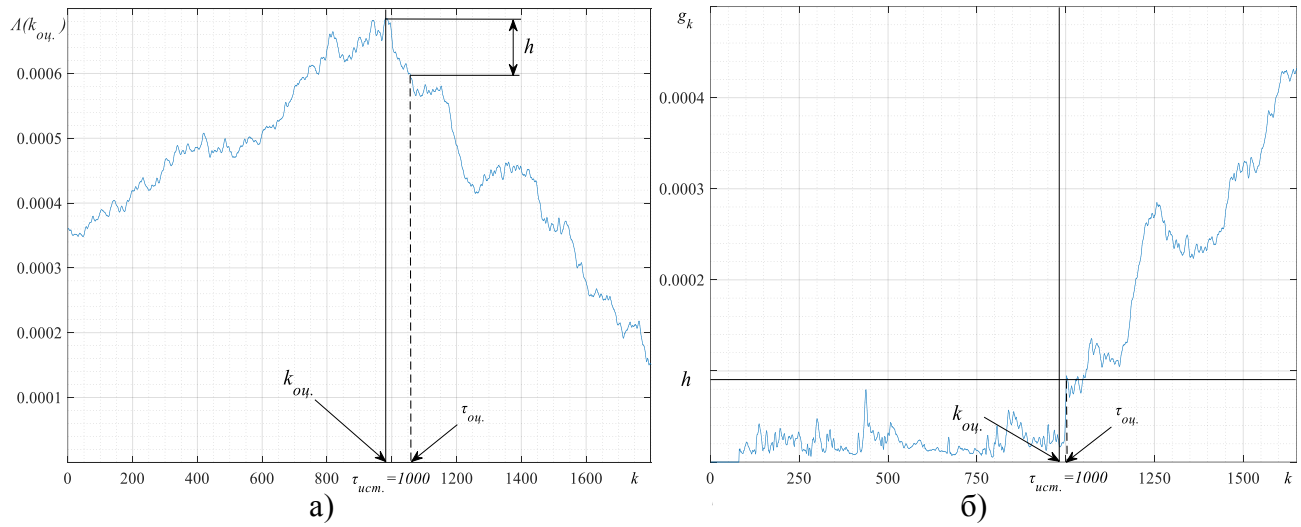

Рисунок 3. (а) Оценка максимального правдоподобия номера точки $k_{\text{оц.}}$; (б) Правило принятия решения о моменте разладки $\tau_{\text{оц.}}$ при оценке координаты $k_{\text{оц.}}$

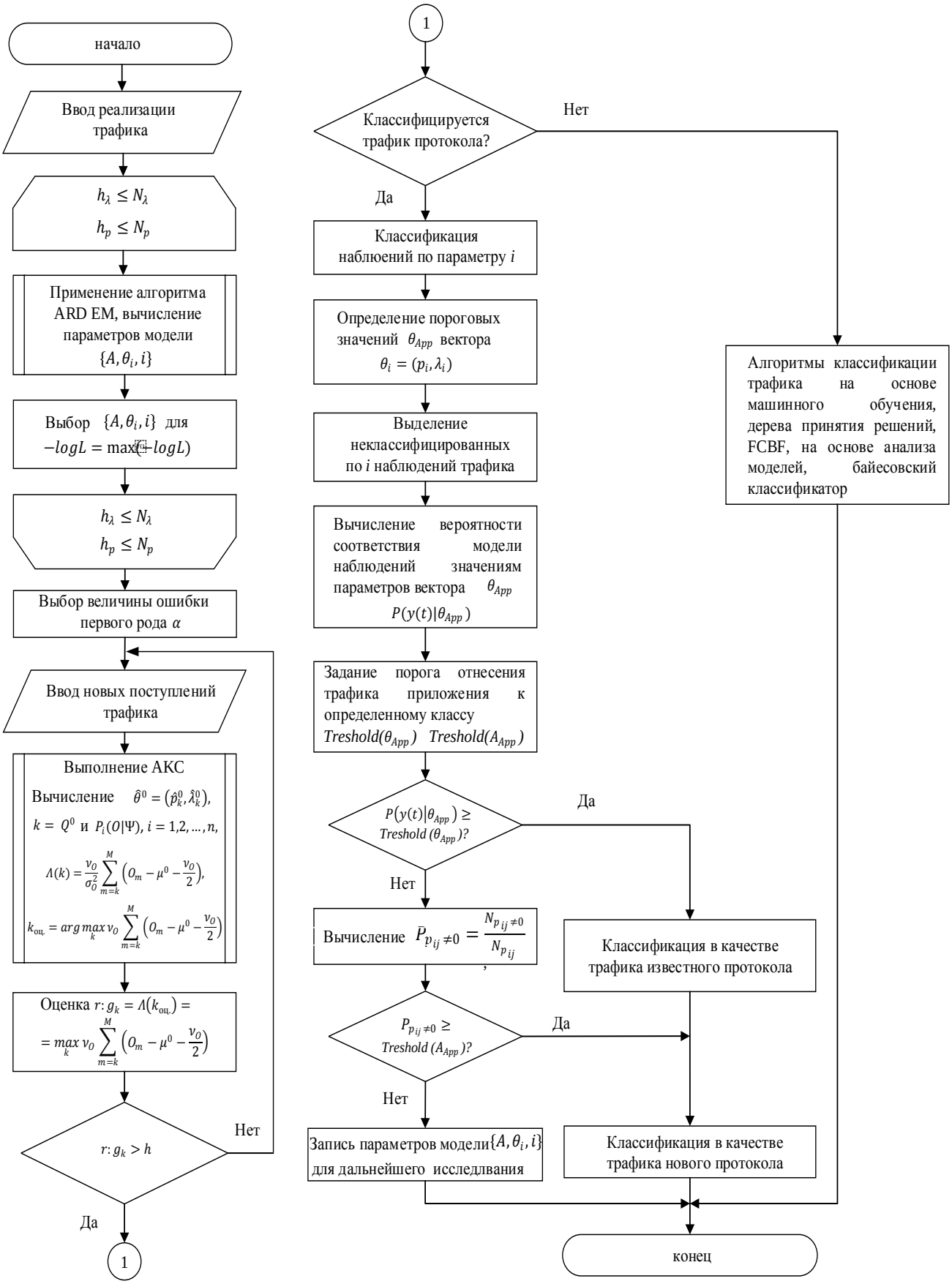

Рисунок 4. Алгоритм идентификации трафика корпоративной телекоммуникационной сети с пакетной коммутацией

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

При выполнении диссертации получены следующие основные результаты:

1. По результатам проведенного анализа состояния и перспективного развития современных вычислительных сетей сформулирована задача идентификации трафика при решении задач управления, контроля и диагностики функционирования корпоративной телекоммуникационной сети с пакетной коммутацией.

2. Произведена оценка существующих методов идентификации трафика корпоративной сети с пакетной коммутацией, на основании которой сформулированы требования к алгоритму идентификации, Выполнение этих требований позволило осуществлять классификацию протоколов сетевого взаимодействия и работать в условиях высокоскоростной передачи данных в реальном времени, что улучшило технические и эксплуатационные характеристики функционирования корпоративной телекоммуникационной сети.

3. По результатам сравнительного анализа существующих моделей сетевого трафика предложена модель, обладающая новизной в части определения числа скрытых состояний.

4. Разработан алгоритм классификации трафика протоколов, используемых в корпоративных телекоммуникационных сетях. Точность проведения классификации достигает 90%, полнота классификации достигает 80%.

5. Проведен анализ влияния загруженности канала передачи информации на параметры модели трафика и исследованы индивидуальные особенности трафика сетевых устройств и пользователей с целью контроля надежности функционирования компьютерных сетей.

6. Решена задача обнаружения изменения свойств трафика телекоммуникационной сети с пакетной коммутацией в реальном времени. Предложен алгоритм последовательного обнаружения момента разладки в наблюдениях трафика сетей передачи данных. Определен минимальный объем обрабатываемых данных для применения указанного алгоритма, а также разработан метод скорейшего обнаружения момента разладки в условиях априорной неопределенности относительно параметров модели трафика.

7. Представлен порядок решения задачи идентификации трафика сети с пакетной коммутацией, применяемый для исследования взаимодействия компьютерных сетей, построенных с использованием различных телекоммуникационных технологий. На основании этого порядка разработан и представлен алгоритм идентификации трафика корпоративной телекоммуникационной сети с пакетной коммутацией.

Основные научные публикации по теме диссертационного исследования

Публикации в изданиях из перечня ВАК России

1. Усовик С.В. Модель трафика вычислительной сети с пакетной коммутацией при априорно неизвестной интенсивности поступления нагрузки. // Известия ОрелГТУ. Информационные системы и технологии. 2010. № 2/58 (585). С. 115-119.

2. Усовик С.В., Воронин А.В. Алгоритм классификации трафика телекоммуникационной сети. // Известия ОрелГТУ. Информационные системы и технологии. 2011. № 1 (63). С. 107-110.

3. Андреев А.М., Усовик С. В. Модель трафика корпоративной телекоммуникационной сети с пакетной коммутацией в задаче кластеризации при условии ограниченного наблюдения. // Инженерный журнал: наука и инновации. – 2012. - № 11. Сайт инженерного журнала: наука и инновации. URL. http://engjournal.ru/articles/485/485.pdf (дата обращения: 10.02.2022).

4. Андреев А.М., Усовик С.В. Статистическое демультиплексирование как задача распознавания сигнала. // Системы высокой доступности. 2016. Т. 12, № 3. С. 39-48.

5. Сюзев В.В., Андреев А.М., Джаммул С.М., Усовик С.В. Анализ методов классификации трафика сетей передачи данных // Динамика сложных систем — XXI век. 2018. № 2. С. 18-28.

6. Андреев А.М., Усовик С.В., Джаммул С.М. Последовательное обнаружение момента разладки наблюдений скрытой марковской цепи при неизвестных параметрах в задаче разделения группового траффика сети с пакетной коммутацией // Наукоемкие технологии. 2018. Т. 19, № 5. С. 12-23.

Статьи и научные труды, опубликованные в других изданиях

7. Андреев А.М., Усовик С. В. Модель трафика корпоративной телекоммуникационной сети с пакетной коммутацией в задаче кластеризации при условии ограниченного наблюдения. // Вестник МГТУ им. Н.Э.Баумана. Серия "Приборостроение". 2012. - Спец. выпуск "Моделирование и идентификация компьютерных систем и сетей". – С. 133-152.