УДК 004.321

А. М. Андреев, С. В. Усовик

МОДЕЛЬ ТРАФИКА КОРПОРАТИВНОЙ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ С ПАКЕТНОЙ КОММУТАЦИЕЙ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ ПРИ УСЛОВИИ ОГРАНИЧЕННОГО НАБЛЮДЕНИЯ

Предложен способ кластеризации параметров трафика сети с пакетной коммутацией, позволяющий упростить задачу разделения трафика по числу абонентов для осуществления контроля на коммутационном узле. Решена задача статистического временного демультиплексирования.

E-mail: 2us@rambler.ru

Ключевые слова: *трафик, телекоммуникационные сети, защита информации, кластеризация, распределение Пуассона, марковская модель.*

При проектировании современных систем передачи информации с пакетной коммутацией, а также при контроле их работоспособности возникает задача контроля трафика на одном из коммутационных узлов: маршрутизаторе, коммутаторе и др. При современном уровне использования администраторами корпоративных телекоммуникационных сетей средств защиты информации, а также средств сетевой защиты решение этой задачи не является тривиальным. Среди основных проблем следует выделить невозможность сопоставления адресной информации, содержащейся в протокольной части передаваемых пакетов, для поиска прямых и обратных каналов. В результате этого невозможно контролировать число абонентов, чей трафик передается по каналу, и типов нагрузки. Вышеперечисленное может оказать влияние на загруженность канала связи, передачу трафика, не предназначенного для данного канала, на невозможность контроля передачи несанкционированного трафика.

В качестве исходных данных для кластеризации выступают наблюдаемые параметры трафика: интервалы между пакетами, длины пакетов, число пакетов, переданных за конкретный период наблюдения. Эти характеристики сетевого трафика остаются неизменными при применении различных средств защиты информации и сокрытия информации об адресатах и применяемых протоколах. Под ограниченностью наблюдения в настоящей работе понимается возможность контроля только одного направления передачи информации. Целесообразность введения такого ограничения обусловлена тем, что даже при прохождении через коммутационный центр трафика другого направления контролируемой корпоративной сети установление этого

факта невозможно из-за отсутствия доступа к протокольной и адресной информации.

Предпосылками к предлагаемой работе стали статьи [1–3]. В них приведены алгоритмы и способы идентификации типа нагрузки, передаваемой в сети Интернет, на основе применения теории марковских, а также скрытых марковских моделей. В частности, были рассмотрены вопросы кластеризации трафика. Достоинством указанных работ служит то, что в них впервые затронута проблема контроля качества обслуживания (QoS) и обнаружения несанкционированного воздействия на сеть передачи данных при использовании пользователями средств туннельного шифрования (IPSec) и сетевых экранов на базе технологии NAT. В работах [1, 3] предложено анализировать интервалы между пакетами и длины пакетов как неизменяемые пользователем характеристики сетевого трафика. Однако в большинстве указанных работ исследования проводились на основе трафика отдельного абонента, т.е. когда входные данные разделены по принадлежности к абоненту. В других работах [4] рассматриваются каналы передачи однородного типа нагрузки (Web, E-mail, FTP). Оба случая являются упрощением и крайне редко встречаются в современных сетях передачи данных с пакетной коммутацией, что делает полученные результаты не применимыми на практике.

Теоретической основой для решения задачи кластеризации трафика телекоммуникационной сети послужил тот факт, что случайный процесс, описывающий поступление пакетов на вход коммутационного устройства не является пуассоновским. Несмотря на то, что процесс поступления заявок от абонента описывается распределением Пуассона, суммарный трафик не удовлетворяет этому распределению. Это утверждение подтверждается практикой. На рис. 1 приведен пример плотности распределения вероятности случайного процесса, описывающего число поступивших пакетов. Реализация трафика получена из магистральной линии, причем пример содержит Web-трафик.

Аппроксимация эмпирической плотности распределения вероятностей с применением критерия Колмогорова–Смирнова [5] подтверждает непуассоновский характер случайного процесса. При этом скорость передачи анализируемого трафика близка к значению пропускной способности канала. Это входит в противоречие с теорией построения очередей [6] и теорией точечных процессов [7].

Рассмотрению этой проблемы посвящено довольно большое число работ [8–13]. В работе [10] приведены подробный обзор существующих моделей сетевого трафика и подробная их классификация. Главным результатом работы является то, что фрактальные модели точнее отражают поведение сетевого трафика. Результаты проектирования вычислительных и телекоммуникационных сетей на основе

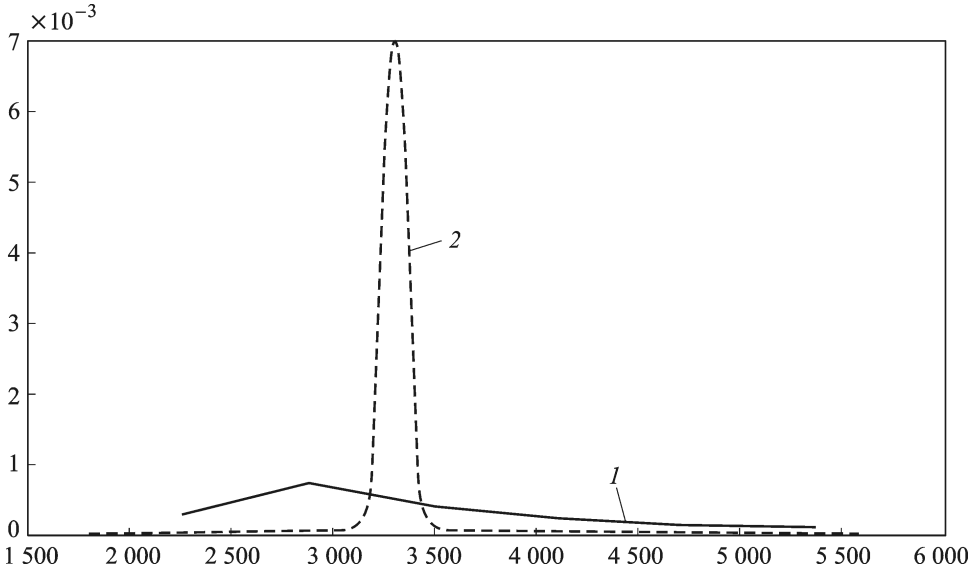

Рис. 1. Эмпирическая (*1*) плотность распределения вероятностей реализации трафика и ее аппроксимация (*2*) распределением Пуассона

фрактальных моделей поведения трафика позволяют поддерживать качество обслуживания на требуемом уровне. Таким образом, алгоритм функционирования протокольной части современных телекоммуникационных сетей с пакетной коммутацией описывается математическим аппаратом теории фракталов.

Для установления наличия эффекта самоподобия в трафике вычислительной сети был проведен эксперимент по наблюдению за трафиком локальной вычислительной сети и изучению ее статистических свойств, в частности периодов поступления пакетов пользователей. Сбор статистики проводился в течение трех дней в различное время. Наблюдались пакеты с трех исходящих IP-адресов: 192.168.1.78, 192.168.1.41 и 192.168.1.50. Составлены временные ряды, содержащие периоды поступления пакетов. Наиболее эффективными методами для оценки самоподобия во временных рядах являются методы анализа R/S-статистики (анализа нормированного размаха) и метод измерения дисперсии на отсчет [10]. Согласно методу анализа нормированного размаха вводится понятие размаха:

$$R(n) = \max_{1 \le j \le n} \Delta_j - \min_{1 \le j \le n} \Delta_j,$$

где

$$\Delta_k = \sum_{i=1}^{k} X_i - k\overline{X}, \quad \forall k = \overline{1, n}$$

— разность между максимальным и минимальным отклонениями.

Для описания изменчивости значений временно́го ряда удобна нормированная безразмерная характеристика

$$\frac{R(n)}{S(n)} = \frac{\max\limits_{1\le j\le n}\Delta_j - \min\limits_{1\le j\le n}}{\frac{1}{n}\sum_{j=1}^{n}[X_j - \overline{X}]^2} = \\ = \frac{\max(0,\Delta_1,\Delta_2,\ldots,\Delta_n) - \min(0,\Delta_1,\Delta_2,\ldots,\Delta_n)}{S(n)}.$$

По результатам проведенных наблюдений вычислены значения коэффициента Херста H, который оценивается по наклону прямой, полученной по графику зависимости $\log\left\{M[]\frac{R(n)}{S(n)}\right\}$ от $\log(n)$. Полученные значения коэффициента Херста для временных рядов, состоящих из периодов поступления пакетов, приведены на рис. 2.

Согласно рис. 2, на основе свойств временны́х рядов, описанных в работе [11], отличия в поведении определенного пользователя наблюдаются на протяжении всего времени наблюдения. Следовательно, наблюдение и анализ периодов поступления пакетов пользователей (хостов) в сети дают устойчивый признак для разделения трафика различных пользователей или фильтрации трафика определенного абонента. Однако эксперимент, проведенный для большого числа абонентов, показывает, что значения коэффициента Херста близки друг к другу. Применение в данном случае байесовского аппарата разделения гипотез [14] приводит к большим ошибкам первого и второго рода.

Главной проблемой применения фрактальной теории для статистического демультиплексирования является невозможность введения вероятностных мер для гипотез в отличие от классического байесовского классификатора [14]. Причиной тому является применяемый ма-

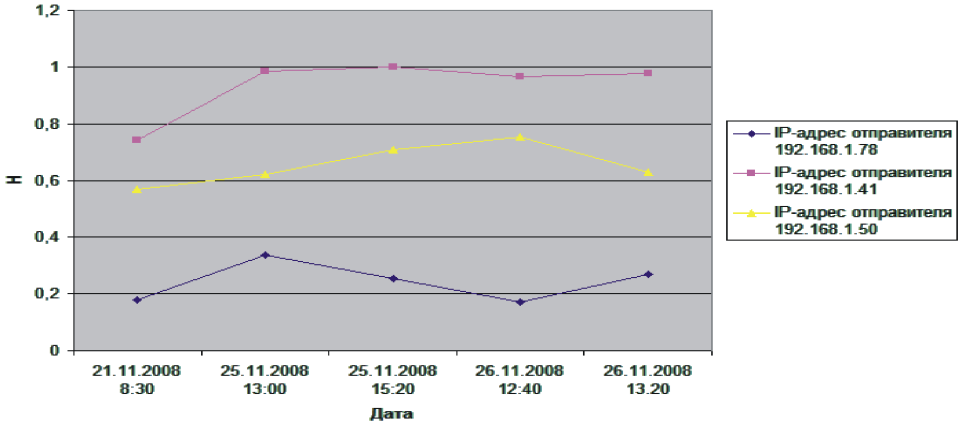

Рис. 2. График значений коэффициента Херста для пакетов наблюдаемых IP-адресов

тематический аппарат, который не использует вероятности для анализа долговременной корреляционной зависимости. В результате по значению коэффициента Херста можно предсказать, будут ли статистические характеристики процесса изменяться в дальнейшем таким образом, как за период наблюдения ($H > 0{,}5$), или в противоположном направлении ($H < 0{,}5$). Математически задача разделения смеси фрактальных распределений остается нерешенной. Решение этой задачи позволит в дальнейшем решить задачу статистического демультиплексирования. Однако в работе [15] совершена попытка с помощью теории мультифракталов определить число абонентов, одновременно участвующих в обмене трафиком.

Предлагается использовать модели, применяющие классический математический аппарат теории вероятностей и адекватно описывающие трафик. Это модели на основе марковских и скрытых марковских процессов. Подобные модели приведены в работах [10, 16, 17]. Скрытая марковская модель с k состояниями $\{s_1, \ldots, s_k\}$ описывается параметрами $\pi_0,\ A,\ B$, где

$$\pi_0 = \{\pi_1, \ldots, \pi_k\}, \quad \pi_k = P(q_0 = i)$$

— вероятность начального состояния;

$$A = \{a_{ij}\},\ 1 \le i, j \le Q, \quad a_{ij} = P(q_{t+1} = i | q_t = j)$$

— матрица переходных состояний;

$$B = \{b_{ij}\},\ 1 \le i \le Q,\ 1 \le j \le M, \quad b_{ij} = P(x_t = i | q_t = j)$$

— вероятность появления наблюдений.

Идея использования скрытых марковских моделей для анализа сетевых процессов состоит в том, чтобы определенная статистическая характеристика трафика зависела от поведения марковской цепи, однако эта марковская цепь скрыта. Предположим, что наблюдения порождены сменой скрытых состояний марковских моделей. Для дальнейшей работы необходимо определить параметры марковской модели, порождающей наблюдения. Попытка раскрыть скрытую марковскую цепь сводится к тому, что каждое состояние s_i соответствует элементу наблюдаемой последовательности x_i. В этом случае число состояний цепи ограничено $N_{\max}$. Исходя из логики работы сетевых приложений, марковская цепь является лево-правой марковской моделью, показанной в работе [16], т.е. $a_{ij} = 0$, если $j \ne i + 1$ и $a_{i,t+1} = 1$. Параметр a_{ij} вычисляется как

$$a_{ij} = \frac{N_{(x_{t+1}=r | x_t=s)}}{N_{(x_t=s)}}, \tag{1}$$

где $N_{(x_{t+1}=r|x_t=s)}$ — число переходов x_i из значения r в значение s, а $N_{(x_t=s)}$ — число значений x_t, равных s. Корреляции между значениями, полученными при помощи формулы (1), определяют близость состояний s в скрытой марковской модели. Параметр b_{ij} определяется

как

$$b_{ij} = \frac{N_{(x_t=j)}}{N_{(x_i)}}, \tag{2}$$

где $N_{(x_t=j)}$ — число x_t в состоянии j, а $N_{(x_i)}$ — общее число x_i. Эта вероятность позволяет судить о близких, в смысле принадлежности к одному скрытому состоянию, значениях x_i.

Далее предполагаем, что каждая последовательность $x_i \in X$ порождена некоторой скрытой марковской моделью H_i. Исходя из этого предположения вычисляется функция правдоподобия $L_{ij} = -\log P(x_i|H_i)$. По максимуму этой функции делается вывод о том, что x_i порождена H_j.

В работе [10] описана модель марковского процесса — это MMPP-модель (пуассоновский процесс, управляемый марковским процессом). Фактически, это модель на основе скрытой марковской модели. Наблюдаемая последовательность x_i представляет собой смесь пуассоновских распределений с интенсивностями $\{\lambda_j\}$. Адекватность этой модели проверяется на речевом трафике, передаваемом посредством пакетов речевого кодека G.703. В этом случае множество состояний интенсивности будет выглядеть как $\lambda_j = \{\lambda_1, \lambda_2\}$, где λ_1 — интенсивность поступления пакетов, когда абонент активен, λ_2 — интенсивность поступления пакетов, когда абонент молчит. В случае смешанного трафика $\lambda_j = \{\lambda_k, \lambda_d\}$, где λ_k — интенсивность поступления голосовых пакетов, λ_d — интенсивность поступления пакетов данных. В модели MMPP [10] данные передаются с постоянной интенсивностью, а для применения ее на практике необходимы априорные данные о типе передаваемого трафика, а также возможность идентификации трафика отдельного абонента, что делает ее непригодной для решения поставленных задач.

В работе [1] описываются модели агрегированного трафика одного типа с применением скрытых марковских моделей. Плотность распределения вероятностей наблюдаемых случайных величин аппроксимируется смесью γ-распределений. Математическая модель представляет собой набор параметров $\Lambda = \{A, g^{(t)}, w^{(t)}, g^{(p)}, w^{(p)}\}$, где A — матрица переходных вероятностей скрытых состояний: $A_{i,j} = P(x_{n+1} = s_j | x_n = s_i)$; $g^{(t)}, w^{(t)}$ — параметры γ-распределения, согласно которому распределены состояния, управляющие интервалами времени между пакетами; $g^{(p)}, w^{(p)}$ — параметры γ-распределения, согласно которому распределены состояния, управляющие длинами пакетов. Смесь γ-распределений выбрана, исходя из того, что подобно смеси нормальных законов распределения ими удобно аппроксимировать произвольное распределение, кроме того, γ-распределение не определено для отрицательных значений случайной величины, что соответствует физической природе интервалов между пакетами и длин

пакетов трафика. Достоинством предложенной модели является ее относительная простота, адекватность физическим процессам, происходящим в телекоммуникационной сети. Модель инвариантна тому, принадлежит ли трафик одному абоненту или является агрегированным трафиком. Однако существенным недостатком является то, что в работе число γ-распределений смеси, аппроксимирующей плотность распределения вероятностей, берется произвольно, без обоснования методики выбора этого числа. Также не приводится сравнения с другими методами аппроксимации плотности распределения наблюдаемых случайных величин. Другим недостатком, снижающим практическое использование алгоритмов идентификации трафика на основе предложенной модели, является то, что анализу должен подвергаться только трафик определенного вида. При этом для решения задач прогнозирования нагрузки модель подходит. Для трафика HTTP и SMTP (наиболее распространенных в современных телекоммуникационных сетях) ошибка прогнозирования составляет 24–40 %.

Работы [1–3, 10] показывают работоспособность моделей, основанных на скрытых марковских моделях. Используем этот результат для описания математической модели агрегированного неоднородного трафика телекоммуникационной сети с пакетной коммутацией. Неоднородность трафика подразумевает наличие пакетов различных прикладных протоколов. При этом также продолжаем накладывать условие ограниченности наблюдения, когда нет физической возможности наблюдать или идентифицировать дуплексные каналы, принадлежащие одному сеансу связи.

Рассмотрим трафик наблюдаемого канала связи как временной ряд, состоящий из числа поступлений пакетов за единицу времени: $X = (x_1, x_2, \ldots, x_n)$ — временной ряд случайных величин x_k размером n, где $x_1, x_2, \ldots, x_n$ — число пакетов, поступивших за единицу времени. Поскольку условия, предъявляемые к трафику в нашем случае, не рассматривались ранее, необходимо проверить соответствие временного ряда X пуассоновскому процессу поступления заявок.

В качестве наблюдаемого возьмем трафик, передающийся с помощью оборудования Cisco и использующий процедуру HDLC в качестве канального протокола. Рассмотрение данного примера обосновано тем, что порядка 80 % трафика телекоммуникационных сетей с пакетной коммутацией в сегменте спутниковой связи построено на оборудовании Cisco, а HDLC-подобные канальные протоколы продолжают активно использоваться в различных видах связи: спутниковой, волоконной. При известной скорости канала упрощается процесс снятия характеристик трафика, поскольку любая временная характеристика может быть получена через подсчет числа переданных байт:

$$t = \frac{N}{V},$$

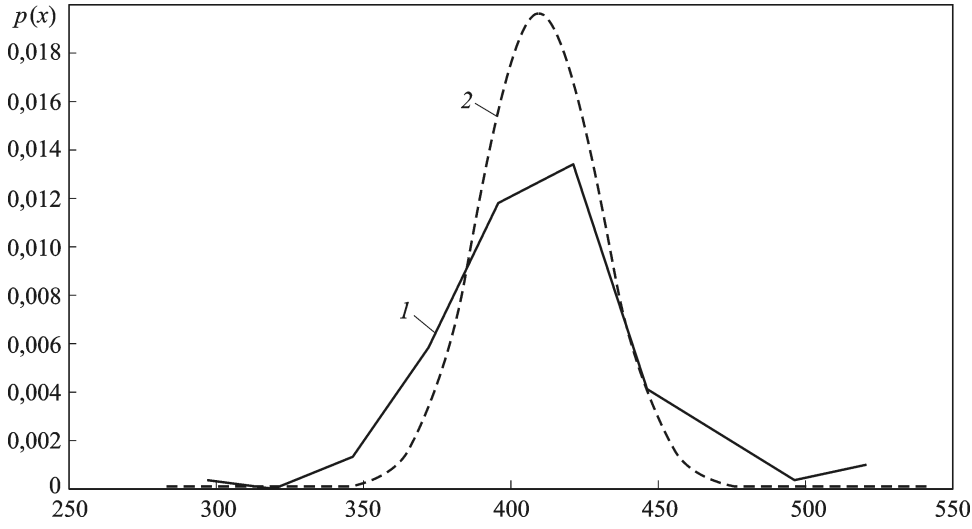

Рис. 3. Аппроксимация наблюдаемой плотности распределения вероятностей распределением Пуассона:
1 — входные данные; *2* — распределение Пуассона

где t — время передачи; N — число байт; V — скорость передачи канала связи.

В рассматриваемом примере случайный процесс является дискретным, поскольку время квантуется на отрезки, необходимые для передачи одного байта.

На рис. 3 представлена плотность распределения вероятностей наблюдаемого процесса поступления пакетов и аппроксимация Пуассоновским процессом.

Распределение Пуассона обладает свойством равенства математического ожидания $\mu(x)$ и дисперсии $v(x)$. Для рассматриваемого примера при использовании критерия Колмогорова–Смирнова и доверительного интервала 95 % имеем $\mu\,(x) = [406{,}92; 413{,}99]$, $v\,(x) = 1284{,}8$. Таким образом, наблюдаемый случайный процесс не соответствует процессу Пуассона. Тест Колмогорова–Смирнова на соответствие нормальному закону распределения, проведенный в среде MatLab, также показал, что наблюдаемое распределение не является нормальным. Проведем аппроксимацию смесью распределений.

Для дискретного случайного процесса суммарная плотность распределения смеси для m компонентов выглядит следующим образом:

$$p\,(x) = \sum_{i=1}^{k} w_i p_i(x),$$

где $p_i, \ldots, p_k$ — функции плотностей распределения вероятностей компонентов смеси; $w_1, \ldots, w_k$ — вероятности соответствующих компонентов смеси. Эту формулу можно представить в следующем виде

для случая, когда наблюдаемый случайный процесс порожден ненаблюдаемыми состояниями:

$$P\left(X=x\right)=\sum_{i=1}^{k}P\left(X=x\mid C=i\right)P(C=i),$$

где $P(X=x)$ — вероятность появления значения X случайной величины x; $P(X=x|C=i)$ — условная вероятность того, что значение X порождено состоянием $C=i$; $P(C=i)$ — вероятность возникновения состояния $C=i$.

Для дальнейшей аппроксимации необходимо провести оценку параметров распределений $p_1,\ldots,p_k$, а также оценить вероятности появления компонентов $w_1,\ldots,w_k$ и число компонентов k. Оценку параметров смеси распределений проведем по методу максимального правдоподобия [8, 9, 15]. В общем случае функция правдоподобия модели смеси m компонентов выглядит следующим образом:

$$L\left(\theta_1,\ldots,\theta_k,w_1,\ldots,w_k\mid x_1,\ldots,x_m\right)=\prod_{j=1}^{m}\sum_{i=1}^{k}w_ip_i\left(x_j,\theta_i\right);$$

здесь $\theta_1,\ldots,\ \theta_k$ — векторы параметров распределений компонентов; $w_1,\ldots,w_k$ — вероятности появления компонентов смеси: $\sum\limits_{i=1}^{k}w_i=1$; $x_1,\ldots,x_m$ — наблюдения размером m. В дальнейшем будем обозначать совокупность всех параметров смеси через $\Xi=\{\Theta,w\}=(\theta_i,w_i)_{i=1}^{k}$.

При условии, что функции распределения компонентов смеси заданы параметрически $p_i(x)=p(x|\theta_i)$, $i=1,\ldots,k$, необходимо решить задачу максимизации функции правдоподобия Ξ_{ML}:

$$\Xi_{ML}=arg\max_{\Xi}p\left(X\mid\Xi\right)\ \ =arg\max_{\Xi}\prod_{j=1}^{m}p(x_j|\Xi).$$

Переходя к логарифму правдоподобия, получаем следующую задачу условной оптимизации:

$$L\left(X,\Xi\right)=\log p\left(X\mid\Xi\right)=\sum_{j=1}^{m}\log\sum_{i=1}^{k}w_ip\left(x_j\mid\theta_i\right)\rightarrow\max_{\Xi}L\left(X,\Xi\right),$$

$$\sum_{i=1}^{k}w_i=1,\quad w_i\geq0,\ \ i=1\ldots k.$$

Приведенный функционал имеет вид “логарифм суммы” и сложен для прямой оптимизации. Наиболее распространенным путем решения этой задачи является применение EM-алгоритма [16, 18, 19].

Известно, что EM-алгоритм для задачи разделения смесей распределения имеет ряд преимуществ перед другими методами оптимизации, такими как проекция градиента и ньютоновские алгоритмы [19].

EM-алгоритм состоит из итерационного повторения двух шагов. На E-шаге вычисляется ожидаемое значение (expectation) вектора скрытых переменных по текущему приближению вектора параметров Θ. На M-шаге решается задача максимизации правдоподобия (maximization) и находится следующее приближение вектора Θ по текущим значениям.

E-шаг (expectation). Найдем $p(x, \theta_i)$ — плотность вероятности того, что объект x получен из i-го компонента смеси. По формуле условной вероятности:

$$p\left(x, \theta_i\right) = p\left(x\right) P\left(\theta_i \mid x\right) = w_i p_i(x).$$

Введем обозначение $g_{ji} \equiv P(\theta_i|x_j)$. Это неизвестная апостериорная вероятность того, что обучающий объект x_j получен из i-го компонента смеси. Возьмем эти величины в качестве скрытых переменных. Обозначим $G = \left(g_{ji}\right)_{m\times k} = (g_1, \ldots, g_i)$, где g_i — i-й столбец матрицы G. Предполагается, что каждый объект может быть сгенерирован одним и только одним компонентом. Согласно формуле полной вероятности отсюда следует условие нормировки для g_{ij}:

$$\sum_{i=1}^{k} g_{ji} = 1 \quad \text{для всех} \quad j = 1, \ldots, l.$$

Зная параметры компонентов w_i, θ_i, легко вычислить g_{ji} по формуле Байеса:

$$g_{ji} = \frac{w_i p_i(x_j)}{\sum\limits_{s=1}^{k} w_s p_s(x_j)} \quad \text{для всех} \quad i,\ j.$$

В этом и заключается E-шаг алгоритма EM.

M-шаг (maximization). Покажем, что знание значений скрытых переменных g_{ji} и принцип максимума правдоподобия приводят к оптимизационной задаче, допускающей эффективное численное решение. Будем максимизировать логарифм правдоподобия

$$Q\left(\Theta\right) = \ln \prod_{j=1}^{m} p\left(x_j\right) = \sum_{j=1}^{m} \ln \sum_{i=1}^{k} w_i p_j\left(x_j\right) \to \max_{\Theta} Q\left(\Theta\right)$$

при ограничении $\sum\limits_{i=1}^{k} w_i = 1$. Запишем лангранжиан этой оптимизаци-

онной задачи:

$$L\left(\Theta; X^m\right) = \sum_{j=1}^{m} \ln\left(\sum_{i=1}^{k} w_i p_i\left(x_j\right)\right) - \lambda\left(\sum_{i=1}^{k} w_i - 1\right).$$

Приравняем нулю производную лагранжиана по w_i:

$$\frac{\partial L}{\partial w_i} = \sum_{j=1}^{m} \frac{p_i(x_j)}{\sum\limits_{s=1}^{k} w_s p_s(x_j)} - \lambda = 0, \quad i = 1, \ldots, k.$$

Умножим левую и правую части на w_i, просуммируем все k равенств и поменяем местами знаки суммирования по i и по j:

$$\sum_{j=1}^{m} \underbrace{\sum_{i=1}^{k} \frac{p_i(x_j)}{\sum_{s=1}^{k} w_s p_s(x_j)}}_{=1} = \lambda \underbrace{\sum_{i=1}^{k} w_i}_{=1},$$

откуда следует $\lambda = m$.

Теперь снова умножим левую и правую части производной лагранжиана на w_i, подставим $\lambda = m$, и, замечая сходство с формулой g_{ji}, получаем выражение весов компонентов через скрытые переменные:

$$w_i = \frac{1}{m} \sum_{j=1}^{m} \frac{w_i p_i(x_j)}{\sum\limits_{s=1}^{k} w_s p_s(x_j)} = \frac{1}{m} \sum_{j=1}^{m} g_{ji}, \quad i = 1, \ldots, k.$$

Легко проверить, что ограничения-неравенства $w_i \geq 0$ будут выполнены на каждой итерации, если они выполнены для начального приближения.

Приравняем нулю производную лагранжиана по θ_i:

$$\frac{\partial L}{\partial \theta_i} = \sum_{j=1}^{m} \frac{w_i}{\sum\limits_{s=1}^{k} w_s p_s\left(x_j\right)} \frac{\partial}{\partial \theta_i} p_i\left(x_j\right) =$$

$$= \sum_{j=1}^{m} \frac{w_i p_i\left(x_j\right)}{\sum\limits_{s=1}^{k} w_s p_s\left(x_j\right)} \frac{\partial}{\partial \theta_i} \ln\ p_i\left(x_j\right) = \sum_{j=1}^{m} g_{ji} \frac{\partial}{\partial \theta_i} \ln\ p_i\left(x_j\right) =$$

$$= \frac{\partial}{\partial \theta_i} \sum_{j=1}^{m} g_{ji} \ln\ p_i\left(x_j\right) = 0, \quad i = 1, \ldots, k.$$

Полученное условие совпадает с необходимым условием максимума в задаче максимизации взвешенного правдоподобия.

Таким образом, М-шаг сводится к вычислению весов компонентов w_i как средних арифметических и оцениванию параметров компонентов θ_i путем решения k независимых оптимизационных задач. Условия сходимости алгоритма EM рассматриваются в работах [9, 16, 18, 19].

Достаточно подробно в литературе описаны аналитические выражения для EM-алгоритма в случае разделения смеси нормальных распределений, т.е., когда в качестве компонентов смеси выбираются следующие:

$$p_i(x) = (2\pi\sigma_i^2)^{-1/2}\exp\left(-\frac{1}{2\sigma_i^2}(x-\mu_i)^2\right).$$

В этом случае максимизация $\Xi = \{\Theta, w\} = (\theta_i, w_i)_{i=1}^k$, где $\theta_i = (\mu_i, \sigma_i)$, на M-шаге выполняется аналитически:

$$w_i = \frac{1}{m}\sum_{j=1}^{m} g_{ji} \quad \text{для всех} \quad i = 1, \ldots, k;$$

$$\mu_i = \frac{1}{mw_i}\sum_{j=1}^{m} g_{ji}x_i \quad \text{для всех} \quad i = 1, \ldots, k;$$

$$\sigma_i^2 = \frac{1}{mw_i}\sum_{j=1}^{m} g_{ji}(x_i - \mu_i)^2 \quad \text{для всех} \quad i = 1, \ldots, k.$$

Отталкиваясь от работоспособности MMPP-модели, необходимо проверить EM-алгоритм для разделения смеси распределений Пуассона, когда в качестве компонентов смеси выступают

$$p_i(x) = \frac{e^{-\lambda_i}\lambda_i^x}{x!}.$$

Пользуясь приведенным описанием E- и M-шагов EM-алгоритма и методом CDLL (complete-data log-likelihood), описанным в [8, 9], для максимизации $\Xi = \{\Theta, w\} = (\theta_i, w_i)_{i=1}^k$, где $\theta_i = \lambda_i$, М-шаг также выполняется аналитически:

$$w_i = \frac{1}{m}\sum_{j=1}^{m} g_{ji} \quad \text{для всех} \quad i = 1, \ldots, k;$$

$$\lambda_i = \frac{\sum_{j=1}^{m} w_j x_i}{\sum_{j=1}^{m} w_j} \quad \text{для всех} \quad i = 1, \ldots, k.$$

Применение EM-алгоритма для восстановления смеси распределений требует задания числа компонентов k. В том случае, если k неиз-

вестно, возникает задача автоматического выбора числа компонентов смеси по данным. Эта задача не может быть решена простым включением k в набор параметров Ξ с дальнейшим поиском параметров по максимуму правдоподобия.

Действительно, чем больше значение k, тем больше значение правдоподобия, так как более гибкая модель может лучше объяснить имеющиеся данные. Выбор числа кластеров является частным случаем проблемы автоматического выбора модели в задачах машинного обучения, заключающейся в наличии ряда параметров, которые не могут быть автоматически определены в рамках классических алгоритмов обучения. Существует довольно много методов определения параметров модели: скользящий контроль, принцип минимальной длины описания (MDL), информационный критерий Акаике, информационный критерий Байеса.

В работе [19] наряду с приведенными методами дано описание алгоритма автоматического определения числа компонентов ARD EM, основанного на методе релевантных векторов. Идея алгоритма состоит в использовании на начальном этапе заведомо избыточного числа компонентов смеси с дальнейшим определением релевантных компонентов с помощью максимизации обоснованности. Подробное описание вывода алгоритма и примера его функционирования приведены в работе [19]. Ключевым моментом являются иные, чем в классическом EM-алгоритме, формулы пересчета весов компонентов:

$$w_i = \frac{\sum\limits_{n=1}^{N} \dfrac{w_i p\left(x_n \mid \theta_i\right)}{\sum\limits_{k=1}^{K} w_k p(x_n|\theta_k)} - w_i^2 \alpha_i}{m - \sum\limits_{k=1}^{K} w_k^2 \alpha_k},$$

где $K = \sqrt{N}$ — начальное число компонентов; α — априорное распределение; $\alpha_i,\ i = 1 \ldots K$, — параметры регуляризации.

В работе [19] алгоритм ARD EM рассматривается на примере смеси нормальных распределений. Из результатов экспериментов кластеризация ARD EM оказывается ближе к истинной, чем у других методов. При этом она практически не уступает по качеству EM-алгоритму с истинным числом кластеров.

Приведем результаты аппроксимации плотности распределения числа пакетов, поступивших за единицу времени, для рассматриваемой реализации трафика.

Сначала проведем аппроксимацию смесью нормальных распределений. Результаты аппроксимации показаны на рис. 4, а также приведены в табл. 1.

На рис. 1 приведены графики аппроксимации смесями двух и трех нормальных распределений, поскольку графики аппроксимации смесями четырех, пяти и шести нормальных распределений подобны графикам, приведенным на рис. 5. Логарифм правдоподобия практически

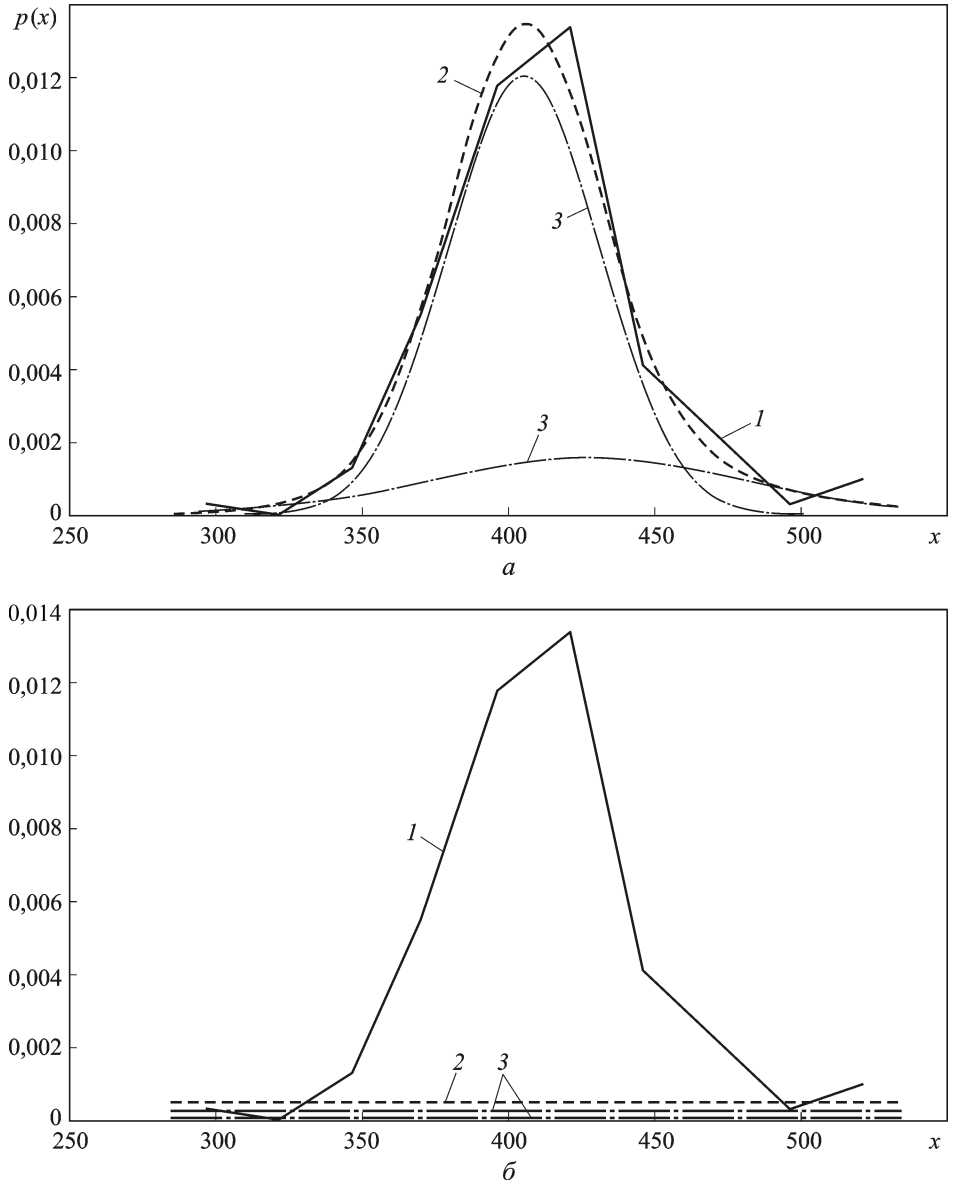

Рис. 4. Аппроксимация плотности распределения смесью двух (*а*) и трех (*б*) нормальных распределений:
1 — эмпирическая ПРВ; *2* — аналитическая ПРВ; *3* — плотности распределения плотности компонентов смеси

Таблица 1

Число компонентов смеси (i)	p_i	μ_i	σ_i	$-\log L$
2	0,7859 0,2141	405,5742 428,3946	26,1263 55,1202	623,4226
3	0,0079 0,8659 0,1262	284 404,4091 459,9493	$1{,}4750\cdot10^{-47}$ 26,5587 39,4351	511,1917
4	0,0079 0,8889 0,0794 0,0238	284 403,9636 462,8248 5206295	$1{,}4750\cdot10^{-47}$ 25,7193 13,0430 9,8475	509,3467
5	0,0079 0,5822 0,2212 0,1649 0,0238	284 390,8118 421,8813 454,7080 520,6316	$1{,}4750\cdot10^{-47}$ 19,8015 7,4503 15,5392 9,8532	506,8277
6	0,0079 0,1492 0,3034 0,3421 0,1737 0,0237	284 365,8511 391,9288 419,6912 453,7003 520,6399	$1{,}4750\cdot10^{-47}$ 10,9004 7,1807 8,4154 16,0128 9,8515	503,5646

не изменяется при увеличении числа компонентов больше 6. При применении алгоритма ARD EM оптимальное число компонентов равно 2.

Далее проведем аппроксимацию смесями распределений Пуассона. Результаты показаны на рис. 5 и приведены в табл. 2.

Логарифм правдоподобия практически не изменяется при добавлении компонентов смеси, когда их число больше 5, причем при добавлении нового компонента смеси, например шестого, параметры распределений становятся неразделимы (λ_3 *и* λ_4).

При сравнении результатов аппроксимации опытной плотности распределения числа пакетов в единицу времени плотностями смесей нормальных и пуассоновских распределений предпочтительным является выбор смеси распределений Пуассона. Главным преимуществом является возможность статистического разделения на кластеры. В случае смеси нормальных распределений, как следует из рис. 4 и табл. 1, компоненты смеси практически неразделимы. В случае смеси пуассоновских распределений при правильном определении числа компонентов возможно осуществить кластеризацию для последующей обработки, в частности для построения матрицы переходных вероятностей скрытой марковской цепи, управляющей процессом поступления трафика. Определение числа компонентов проводится согласно алгоритму ARD EM, а также по изменению значений правдоподобия.

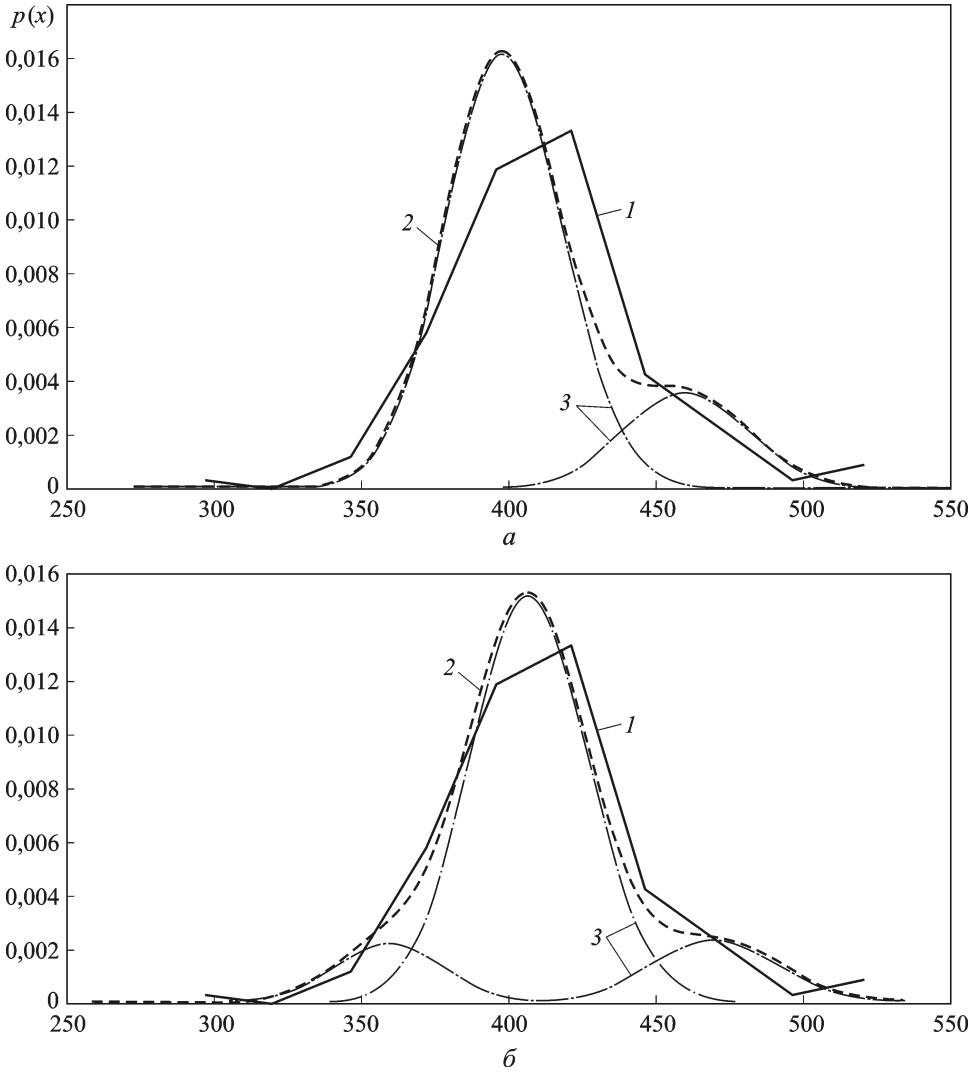

Рис. 5 (начало). Аппроксимация плотности распределения смесью двух (*а*), трех (*б*), четырех (*в*), пяти (*г*) и шести (*д*) распределений Пуассона (*1–3* — обозначения см. рис. 4)

Таким образом, в качестве модели трафика корпоративной телекоммуникационной сети с пакетной коммутацией предлагается математическая модель, представляющая собой набор параметров $\{A, \theta_i, i\}$. В этой модели 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 компонента смеси, λ_i — интенсивность распределения Пуассона (компонента смеси); i — число компонентов смеси. Отличие данной модели трафика от описанных ранее заключается в том, что предложено ввести число компонентов смеси в виде отдельного параметра благодаря возможности получить параметр i аналитически.

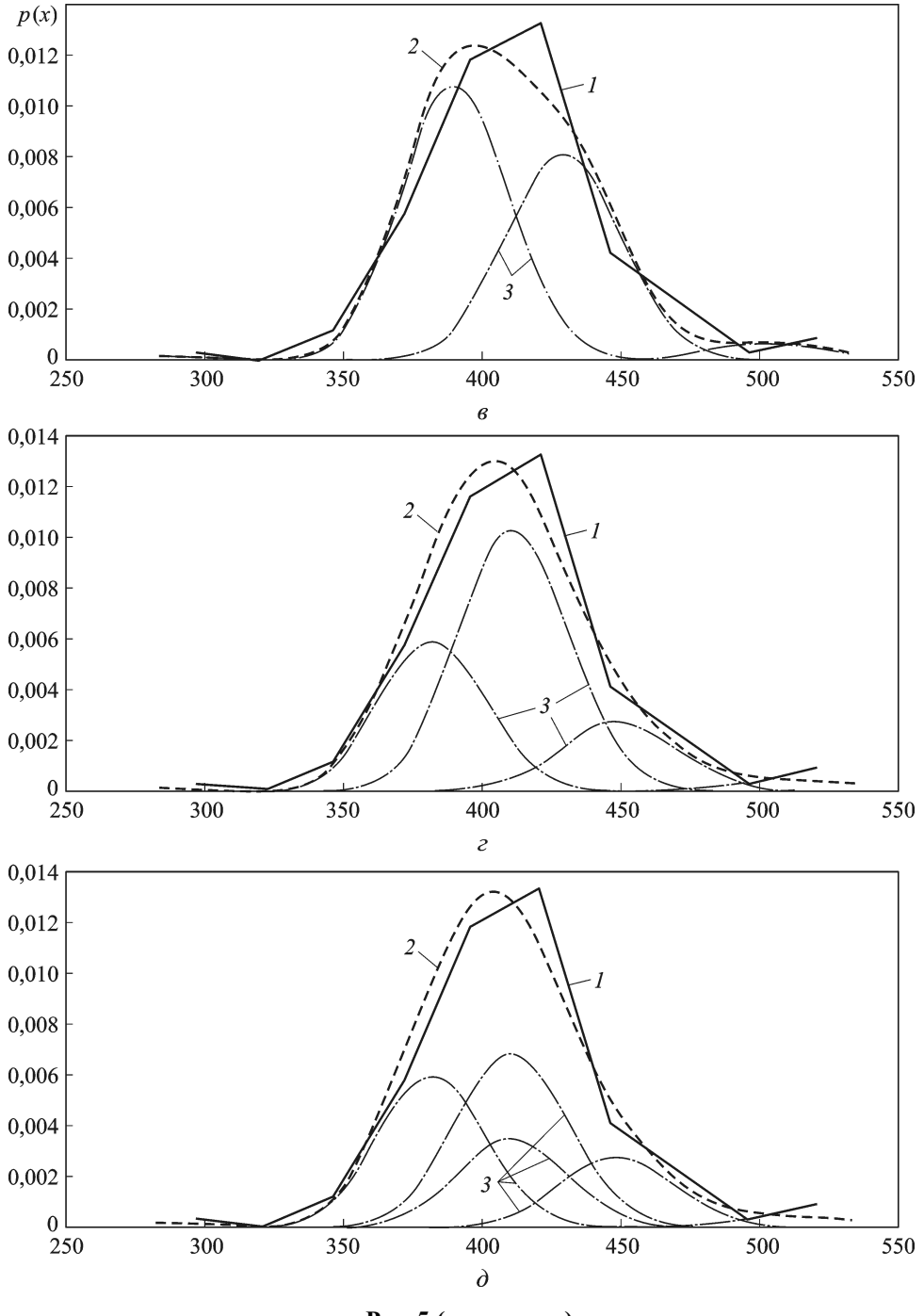

Рис. 5 (окончание)

Число компонентов смеси (i)	p_i	λ_i	$-\log L$
2	0,8079 0,1921	398,4932 460,7815	641,9453
3	0,1042 0,7576 0,1283	359,4410 407,2063 471,3801	631,2463
4	0,0079 0,5334 0,4198 0,0388	284,0454 390,3418 429,7527 504,1384	621,8201
5	0,0079 0,2923 0,5254 0,1468 0,0276	284,0353 382,3691 411,6022 449,7279 513,6957	620,8579
6	0,0079 0,2923 0,1774 0,3480 0,1468 0,0276	284,0353 382,3691 411,6022 411,6022 449,7279 513,6957	620,8579

В качестве компонента смеси предлагается использовать распределение Пуассона (в отличие от [1]), что облегчает процесс построения модели благодаря минимальному числу обрабатываемых параметров. Для смеси распределений Пуассона ускоряется работа EM-алгоритма. Как показал опыт, возможен процесс кластеризации в отличие от смеси нормальных распределений. Благодаря этому модель может быть использована в алгоритмах, применяющих кластеризацию для идентификации вида трафика, для ускорения процесса обработки. Преимуществом перед фрактальными моделями является меньшая вычислительная сложность и возможность применения в алгоритмах на основе предложенной модели классического математического аппарата теории вероятностей и теории принятия решений, которые являются хорошо изученными. Модель также позволяет дополнить модель MMPP [10] тем, что трафик неголосовой нагрузки имеет изменяемую интенсивность.

Предложенная модель характеризует конкретную корпоративную телекоммуникационную сеть, поскольку параметры θ_i и i будут отличаться для различных сетей. Следовательно, это является предпосылкой для разделения сетей, обнаружения процесса смены передачи трафика отличающихся сетей. В работах [17, 20] предложены алго-

ритмы и методы обнаружения момента смены состояний случайного процесса, а также обнаружения момента времени скачкообразного изменения состояния случайного процесса. В этих алгоритмах предложенная модель применима для определения изменяемых структурных параметров при скачкообразном изменении свойств случайного процесса. В алгоритмах [17, 20] ищется момент изменения параметров $\{A, \theta_i, i\}$. Эти параметры могут использоваться для получения эталонов в процессе обучения алгоритмов, предложенных в [17].

Таким образом, предложена новая модель трафика корпоративной телекоммуникационной сети с пакетной коммутацией, обладающая преимуществом перед существующими моделями и находящая применение в существующих алгоритмах кластеризации и идентификации трафика, а также алгоритмах статистического разделения (демультиплексирования).

СПИСОК ЛИТЕРАТУРЫ

1. Dianotti A., Pescape A., Rossi P. S., Palmieri F., Ventre G. Internet traffic modeling by means of hidden Markov models / Computer Networks 52(2008), 2645–2662. www.elsevier.com/locate/comnet
2. Wright C., Monrose F., Masson G.: HMM profiles for network traffic classification (extended abstract) // Proc. of Workshop on Visualization and Data Mining for Computer Security (VizSEC/DMSEC), Fairfax, VA, USA (2004), 9–15.
3. Dainotti A., de Donato W., Pescapè A., Rossi P. S.: Classification of network traffic via packet-level hidden Markov models // Proc. of IEEE Global Telecommunications Conference (GLOBECOM) 2008, New Orleans, LA, USA (2008).
4. Mah B. A. An empirical model of HTTP network traffic // INFOCOM'97. – Vol. 2. – P. 592–600, Apr. 1997.
5. Бронштейн И. Н., Семендяев К. А. Справочник по математике для инженеров и учащихся втузов. – М.: Наука, 1986. – 544 с.
6. Овчаров Л. А. Прикладные задачи теории массового обслуживания. – М.: Машиностроение, 1969. – 324 с.
7. Клейнрок Л. Теория массового обслуживания / Пер. с англ. И.И. Грушко; Под ред. В.И. Нейман. – М.: Машиностроение, 1979. – 432 с.
8. Zucchini W. and MacDonald I. L. Hidden Markov models for time series: An introduction using R. Chapman & Hall (CRC Press), 2009.
9. MacDonald I. L. and Zucchini W. Hidden Markov and other models for discrete-valued time series // London: Chapman and Hall, 1997.
10. Шелухин О. И., Тенякишев А. М., Осин А. В. Фрактальные процессы в телекоммуникациях. Монография / Под ред. О.И. Шелухина. – М.: Радиотехника, 2003. – 480 с.
11. Федер Е. Фракталы: Пер. с англ. – М.: Мир, 1991. – 254 с.
12. Lane T. Hidden Markov models for human/computer interface modeling // Proc. of the IJCAI-99 Workshop on Learning about Users, pp. 35–44. International Joint Conferences on Artifficial Intelligence, August 1999.
13. Crotti M., Dusi M., Gringoli F., Salgarelli L. Traffic classification through Simple statistical fingerprinting / ACM SIGCOMM Computer Communication Review. Vol. 37. No. 1, January 2007.

14. Г м у р м а н В. Е. Теория вероятностей и математическая статистика: Учеб. пособие для вузов. – М.: Высш. шк., 2003. – 478 с.
15. У р ь е в Г. А., Ш е л у х и н О. И., О с и н А. В. Результаты экспериментальных исследований сетевого трафика телекоммуникационной сети // Теоретические и прикладные проблемы сервиса. – 2005. – № 1–2 (14–15). – С. 38–49.
16. R a b i n e r L. R. A tutorial on hidden Markov models and selected applications in speech recognition // Procs. IEEE. – Vol. 77. No. 2. – P. 257–285, Feb. 1989.
17. М о т т л ь В. В., М у ч н и к И. Б. Скрытые марковские модели в структурном анализе сигналов. – М.: Физматлит, 1999. – 352 с.
18. D e m p s t e r A. P., L a i r d N. M. and R u b i n D. B. Maximum likelihood from incomplete data via the EM algorithm // J. of the Royal Statistical Society. Series B (Methodological), 39(1), 1–38, 1977.
19. В е т р о в Д. П., К р о п о т о в Д. А., О с о к и н А. А. Автоматическое определение количества компонент в EM-алгоритме восстановления смеси нормальных распределений // Ж. вычисл. матем. и матем. физ. – 2010. – Т. 50, № 4. – С. 1–14.
20. Ж и г л я в с к и й А. А., К р а с к о в с к и й А. Е. Обнаружение разладки случайных процессов в задачах радиотехники. – Л.: Изд-во ЛГУ, 1988. – 224 с.

Статья поступила в редакцию 15.12.2011