УДК 004.7.052:004.414.2

Якимчук Н.М., Селепина Й.Р., *к.т.н.,* **Євсюк М.М.**, *к.т.н.*
Луцький національний технічний університет

МЕТОДИ ІМПУТАЦІЇ ДАНИХ В СТАТИСТИЧНОМУ АНАЛІЗІ ПАРАМЕТРІВ ТЕЛЕКОМУНІКАЦІЙНИХ МЕРЕЖ

Розглядаються питання оперативної діагностики особливостей телекомунікаційних мереж, що важливі для підвищення якості передачі інформації. Проаналізовано параметри мережі для вирішення завдань мережевого адміністрування, моніторингу мережевого устаткування, виявлення аномальної поведінки чи збоїв в роботі системи. Розглядаються способи отримання статистичних оцінок параметрів мережі, засновані на використанні методів максимізації числа комплектних елементів. Із застосуванням методів регресійно-кореляційного аналізу визначається характер зміни інтенсивності мережевого трафіку, статистичних параметрів трафіку, зростання інтенсивності використання окремих елементів мережевого обладнання.

Ключові слова: *мультисервісна мережа, регресійно-кореляційний аналіз, множинна регресія.*

Постановка проблеми. Телекомунікаційні мережі відносяться до класу складних систем, яким властивий ряд характеристик, основними з яких є:

- різнорідність складових елементів, кожен з яких вирішує свою часткову задачу в рамках єдиної мети функціонування всієї системи;
- складність взаємозв'язків між елементами системи і параметрами, що їх описують;
- багатоплановість вирішення завдань;
- випадковий характер процесів, що протікають в системі;
- багатопараметричний опис системи;
- залежність якості функціонування системи від множини різнорідних фактів, тощо.

Оскільки пропускна здатність і складність мереж постійно збільшуються, вони оснащуються десятками або навіть сотнями пристроїв доступу до мережевого трафіку, які забезпечують його повний контроль. Стрімкий нелінійний ріст потоку даних в результаті вільного доступу до мереж і швидкого розвитку цифрових послуг і додатків, вимагатиме все складнішої інфраструктури, що тягне за собою величезні фінансові затрати фірм-операторів телекомунікаційних послуг. Скоротити затрати можна звернувши увагу на досягнення максимально ефективного розподілу потоків даних та рівномірного завантаження існуючих каналів і пристроїв комутації. Наявність в мережі великого числа таких пристроїв (наприклад, комутаторів і «інтелектуальних» розгалужувачів) не тільки створює хороші можливості моніторингу мережевого трафіку, а й породжує проблему управління всіма цими пристроями.

Задача підвищення якості передачі інформації пов’язана, перш за все, з процедурою оперативної діагностики мережі. В новому інфокомунікаційному середовищі кінцевою послугою, як правило, стає не послуга зв’язку, а контент, що часто передається в мережу сторонніми провайдерами, параметри якого часто оператору невідомі. Проте, поряд з інформаційними даними мережа передає значний об’єм сигнального трафіку, що містить в собі відомості практично про всі аспекти роботи мережі зв’язку, мережеві служби, запити абонентів і надані послуги. Інтелектуальна обробка цих даних дозволить виокремити зміни інтенсивності трафіку, наявність і характер збоїв, параметри інтенсивності використання окремих частин мережі, тощо.

Аналіз літературних джерел. Параметричне діагностування телекомунікаційних систем сьогодні розглядається спрощено, шляхом неперервного чи періодичного спостереження за параметрами трафіку з використанням аналізаторів протоколів, що пояснюється простотою, наочністю і оперативністю контролю [4]. Прийнято вважати, що якщо результати не виходять за встановлені межі, то мережа працездатна. Недоліками такого підходу є неможливість виявити проблеми передачі трафіку на ранніх стадіях їх виникнення, коли параметри ще не досягли граничних значень та недоступність інформації від віддалених вузлів системи, безпосередньо не зв’язаних з контрольованим.

Для розробки та застосування статистичних методів мережевого моніторингу та аналізу, перш за все, необхідно побудувати математичні моделі мережевого трафіку [3,7]. Очевидно, це є задачею математичної статистики [5,8].

На етапі моніторингу виконується процедура збору первинних даних про роботу мережі, статистика про кількість циркулюючих в мережі пакетів, про стан портів концентраторів, комутаторів і маршрутизаторів, тощо. Етап аналізу являє собою складніший і інтелектуальний процес осмислення зібраної в процесі моніторингу інформації, з використанням спеціальних інструментальних засобів і методик обробки часових рядів.

Нові можливості дає нам використання факторного аналізу сигнального трафіку, кореляційний аналіз його окремих складових дає можливість прогнозувати появу змін в системі та успішно вирішувати задачу мережевого адміністрування. Кореляційні зв'язки між факторами і параметрами відображають основну діагностичну інформацію по трафіку, що проходить через той чи інший інтерфейс мережевого пристрою.

Виклад основного матеріалу.

До основних показників якості телекомунікаційної мережі відносять:

- пропускну спроможність мережі - інтегральний показник, що характеризує об'єм інформації, що пропускається,
- реакція на характеристики профілю трафіку - для визначення цієї характеристики мережа моделюється як чорний ящик, розглядається реакція на зміну навантаження на мережу,
- кількість втрачених пакетів - для TCP мережі 1-5% втрачених пакетів, згідно з експертними оцінками, знаходиться в межах норми, 40% втрачених пакетів - граничне значення, при якому мережа практично не працює,
- час доставки - вимірюють часом подвійного ходу (у прямому і зворотному напрямі), цей показник фіксується з використанням програми PING,
- нерівномірність доставки пакетів - ця характеристика впливає на роботу окремих додатків, наприклад, передачі аудіопотоку чи відеоконференції або в пакетній телефонії.

Технічні показники функціонування мережі, як правило, представляються таблицями статистичних даних:

$$\begin{pmatrix} y(1) & y(2) & \cdots & y(i) & \cdots & y(N) \\ x_1(1) & x_1(2) & \cdots & x_1(i) & \cdots & x_1(N) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_j(1) & x_j(2) & \cdots & x_j(i) & \cdots & x_j(N) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_k(1) & x_k(2) & \cdots & x_k(i) & \cdots & x_k(N) \end{pmatrix}.$$

Статистичні дані представляють собою вибірку деякої реалізації значень випадкових величин:

– i -а реалізація чисельного значення результату y_i, $i-1,2,\ldots,N$;

– j -а реалізація чисельного значення j -го фактора x_j, $j=1,2,\ldots,N$.

Загальне призначення множинної регресії полягає в аналізі зв'язку між кількома незалежними змінними (факторами - рядки матриці спостережень $\mathbf{X}$) і залежною змінною (елементи вектора відгуків $\bar{\mathbf{y}}$).[6]. Використання статистичних даних дозволяє домагатися оптимальних результатів, керуючи величинами факторів, або прогнозувати можливу величину результату при сформованих значеннях факторів.

Між випадковою величиною результату і випадковою величиною фактора є стохастична (випадкова) залежність, тобто існує кореляційна залежність.

У загальному випадку, процедура побудови множинної регресії полягає в оцінюванні параметрів лінійного рівняння.

$$E\begin{bmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{pmatrix}\end{bmatrix} = \begin{pmatrix} E[y_1] \\ E[y_2] \\ E[y_3] \\ \vdots \\ E[y_m] \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1n} \\ 1 & x_{21} & x_{22} & \ldots & x_{2n} \\ 1 & x_{31} & x_{32} & \ldots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \ldots & x_{mn} \end{pmatrix} \cdot \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix},$$

або те ж саме в компактному вигляді:

$$E[\vec{y}] = \mathbf{X}\vec{\theta}.$$

Регресійні коефіцієнти представляють незалежні вклади кожної незалежної змінної в передбачення залежної змінної.

Обов'язковим етапом при проведенні інтелектуального аналізу даних є підготовка і очистка даних, що можуть подаватися в різних форматах, містити такі помилки узгодження як неправильні дані, екстремальні викиди чи відсутні дані, сюди відноситься виявлення прихованих залежностей між факторами, що можуть непередбаченим чином впливати на модель. На цьому етапі найбільш перспективним є застосування методів імпутації (відновлення) пропущених значень.

Імпутація - це процедура оцінки невідомих значень показників на основі наявних даних, результатом якої є отримання комплектної множини даних. Методи імпутації можна застосувати й до екстремальних значень параметрів (викидів), що виникають внаслідок появи помилок при отриманні даних і повинні бути видалені. Альтернативами множинної імпутації є методи повного видалення, часткового видалення, заміни середнім та метод максимізації очікування.

- Метод повного виключення полягає в тому, що всі спостереження, де є пропущені значення виключаються з обчислень. Очевидно, що дані методи значно зменшують об'єм вибірки і не дозволяють побудувати адекватну модель системи, вони можуть застосовуватися, коли доля відсутніх значень надзвичайно мала (до 1%) чи умовно допустима 1-5%).З огляду на це метод видалення доцільно використати лише для змінних, більшість елементів яких невідомо для того щоб максимізувати число комплектних елементів вибірки.
- Метод попарного виключення є різновидом попереднього – аналізовані змінні порівнюються попарно, для кожної пари аналізуються тільки ті спостереження, де немає пропущених значень змінних цієї пари.
- Метод заміни середнім значенням – всі пропуски в матриці даних заповнюються значеннями, що дорівнюють середньому арифметичному по даній змінній. Такий метод використовує припущення, що змінна, для якої здійснюється заміна нормально розподілена і крім того володіє низькою дисперсією, тобто імовірність появи середнього значення даного фактора вища ніж будь-якого іншого. Різновидом методу заміни середнім є заміна середнім по найближчих k-сусідах.

Проте більшість процесів в телекомунікаційних системах, що відносяться до класу систем масового обслуговування, описуються як простий потік подій розподілений за законом Пуассона: імовірність того, що за час τ відбудеться рівно m подій рівна:

$$P(X = m) = \frac{(\lambda\tau)^m}{m!} e^{-\lambda\tau}.$$

Даний розподіл володіє вираженою асиметрією, тобто масив буде містити відносно мало значень рівних чи близьких середньому.

- Метод максимізації очікування здійснюється з використанням інформації про вид закону розподілу даної змінної. Спираючись на статистичний аналіз попередніх даних розраховується величина, імовірність появи якої в даній точці спостереження була б максимальною, тобто визначається умовне очікування пропущених даних, після чого підстановки отриманих результатів. Недоліком даного способу є можливість отримання зміщених оцінок, коли модель «підганяється» під заданий закон розподілу, який може бути помилковим.

Цих недоліків позбавлений метод множинної імпутації, хоча він вимагає значно більшої тривалості процедури підготовки даних. На відміну від методу підстановки середнього множинна імпутація не призначена для підстановки конкретних значень в базу даних. Замість цього будується конкретна регресійна модель використовуючи (підбираючи) на місці пропусків ті дані, які давали б найкращі показники передбачення даної моделі. Цей процес здійснюється циклічно в кілька етапів, на початковому етапі на місці пропущених даних встановлюються довільні параметри (частіше середні значення), створюється регресійна модель, оцінюється відхилення прогнозованих значень вихідних параметрів моделі від спостережуваних даних. На наступних етапах здійснюється зміна значень (зсув) імпутованих даних і перерахунок параметрів регресійної моделі в сторону покращення показників прогнозування моделі до досягнення необхідної точності. Для проведення цієї процедури основною умовою є початкове розбиття всіх вхідних даних на дві підмножини: експериментальну, по даних якої будується модель і тестову – з використанням якої здійснюється оцінка якості прогнозування моделі.

Для відбору остаточного рівняння регресії зазвичай використовують два протилежних критерії. *По-перше,* щоб зробити рівняння корисним для передбачення, ми повинні прагнути включити в модель по можливості більше незалежних змінних з тим, щоб можна було більш

надійно визначити прогнозовані величини. *По-друге,* через витрати, пов'язані з отриманням великої кількості інформації і її подальшою перевіркою, необхідно прагнути, щоб рівняння включало якнайменше незалежних змінних.

Компроміс між цими критеріями може бути досягнутий за рахунок вибору "найкращого" рівняння, що включає оптимальну кількість незалежних змінних. В роботі для пошуку "найкращого" рівняння регресії застосований крокoвий метод (покрокова регресія), коли незалежні змінні (регресори) одна за одною включаються в підмножину згідно попередньо заданого критерію. У той же час деяка змінна може бути замінена іншою змінною, яка не входить в набір, або видалена з нього. Процедура визначення числа регресорів називається правилом зупинки – включення або виключення змінної X_j при визначенні F-критерію, що служить статистикою для перевірки того, що змінна X_j значимо покращує чи не покращує якість передбачення [8].

Проведені таким чином процедури підготовки даних підвищуються якість вхідної інформації і дозволяють розкласти часовий ряд на 2 адитивних компоненти: детерміновану частину (пов’язану з сезонним і добовим коливанням завантаження мережі) з явно вираженими періодами максимального і мінімального завантаження і випадкову компоненту. Такий поділ полегшить спектральний аналіз даних, обчислення тренду та оцінку частоти появи викидів завантаження каналу. Кореляція подій дозволяє виокремити значимі події, визначити вагу впливу кожного параметра на результативний чинник. Для оцінки появи аномалій здійснюється порівняння поточних характеристик мережі з характеристиками прийнятими за нормальний профіль роботи з застосуванням методів робастної статистики (М-оцінки).

Результатами діагностики мережі є визначення нормального профілю роботи мережі і виявлення і передбачення відхилень від нього для відповідної зміни конфігурації системи. Для прикладу: збільшення кількості викидів – короткочасне підвищення завантаженості каналу, сигналізує про появу перешкод, шумів на лінії, що вимагають повторної передачі пошкодженої інформації, в комп’ютерних мережах сигналізує про появу вірусів, паразитного трафіку. Зміна спектральної характеристики ряду пов’язана зі зміною коливання добового завантаження мережі, що вказує на зміну режимів роботи користувачів мережі. Зміна тренду показує ріст завантаженості мережі, що може призвести до перевищення пропускної здатності. Стрибкоподібне збільшення середнього значення завантаженості сигналізує, що мережа працює в режимі граничного завантаження.

Висновки. Оперативно здійснювати ефективну оцінку роботи всіх компонентів такої складної системи як телекомунікаційні мережі з урахуванням їх взаємозв'язків і взаємовпливу можливо лише з використанням статистичних методів аналізу і обробки оперативної інформації. Неефективний розподіл ресурсів, недосконала організація взаємодії апаратури обмежують пропускну спроможність мережі, погіршують якість її функціонування. Регресійно-кореляційний аналіз зв’язків між факторами і параметрами роботи комунікаційного обладнання відображають основну діагностичну інформацію по трафіку, що проходить через той чи інший інтерфейс мережевого пристрою, дозволяє прогнозувати зміни інтенсивності робочого навантаження, передбачати і сигналізувати про небезпеку відхилення від нормального профілю роботи мережі.

Складністю проведення статистичного аналізу є необхідність врахування і обробки великої кількості інформації, втрата частини якої внаслідок технічних чи організаційних перешкод чи збоїв веде до появи зміщених статистичних оцінок, погіршення якості моделей. Уникнути цього допоможе застосування методу множинної імпутації даних, що дозволить очистити сигнальну інформацію від випадкових викидів, відновити втрачені дані, уникнути надлишкових, малоінформативних параметрів. що полегшить спектральний аналіз даних, обчислення тренду та оцінку частоти появи викидів завантаження каналу.

Ін формаційні джерела

1. Tanenbaum A. S. Computer networks, 5th ed. / Andrew S. Tanenbaum, David J. Wetherall. – Prentice Hall, Cloth, 2011. – 960 p.

2. Stallings W. Foundations of modern networking: SDN, NFV, QoE, IoT, and Cloud. – Pearson Education, Inc., Old Tappan, New Jersey, 2016. – 538 p.

3. Виноградов Н. А. Анализ потенциальных характеристик устройств коммутации и управления сетями новых поколений / Н. А. Виноградов // Зв’язок. – 2004. – №4. – С. 10-17.

4. Лесная Н. Н. Сравнительный анализ методов оценки характеристик интеллектуальной сети // Наукові записки Українського науково-дослідного інституту зв’язку.

– 2009. – №2(10). – С. 97-102.

5. Афифи А. Статистический анализ: Подход с использованием ЭВМ / А. Афифи, С. Эйзен. – Москва: Мир, 1982. – 488 с.

6. Мостеллер Ф. Анализ данных и регрессия: вып. 1 / Ф. Мостеллер, Дж. Тьюки. – Москва: Финансы и статистика, 1982. – 317 с.

7. Торошанко Я. І. Задачі моніторингу та аналізу параметрів телекомунікаційних мереж / Я. І. Торошанко, А. О. Булаковська, М. С. Височіненко, В. С. Шматко // Телекомунікаційні та інформаційні технології. – 2014. – №3. – С. 62-69.

8. Барабаш Ю. Л. Вопросы статистической теории распознавания / Ю. Л. Барабаш, Б. В. Варский, В. Т. Зиновьев, В. С. Кириченко, В. Ф. Сапегин. – Москва : Советское радио, 1967. – 400 с.

Якимчук Н. Н., Селепина Й.Р., *к.т.н.,* **Евсюк М.М.,** *к.т.н.*
Луцкий национальный технический университет

МЕТОДЫ ИІМПУТАЦИИ ДАННЫХ В СТАТИСТИЧЕСКОМ АНАЛИЗЕ ПАРАМЕТРОВ ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЕЙ

Рассматривается вопрос оперативной диагностики особенностей телекоммуникационных сетей, который является существенным для повышения качества передачи информации. Проанализированы параметры сети для решения заданий сетевого администрирования, мониторинга сетевого оборудования, выявления аномального поведения или сбоев в работе системы. Рассматриваются способы получения статистических оценок параметров сети, основанные на использовании методов максимизации числа комплектных элементов. С применением методов регрессионно-корреляционного анализа определяется характер изменения интенсивности сетевого траффика, статистических параметров траффика, роста интенсивности использования отдельных элементов сетевого оборудования.

***Ключевые слова**: мультисервисная сеть, регрессионно-корреляционный анализ, множественная регрессия.*

N. Yakymchuk, J. Selepyna, M. Yevsyk
Lutsk National Technical University

METHODS OF DATA IMPUTATION IN STATISTICAL ANALYSIS OF TELECOMMUNICATION NETWORKS PARAMETERS

The questions of operative diagnostics of features in telecommunication networks are studied, which are being important for improving the quality of information transfer. To solve the tasks of network administration, monitoring of network equipment, and reveal system abnormal behavior or network disturbance the network state is analyzed. The ways of receipt of statistical estimations of network parameters, based on the methods of maximizing of the number of complete sample elements are considered. The type of network disturbance the network traffic intensity, changes in stationary character of traffic, excessive increase in intensive use of network parts are revealed with the use of the regression-correlation analysis methods.

***Keywords**: multiservice network, regression-correlation analysis, multiple regression.*