УДК 004.7.052:004.414.2

Н.М. ЯКИМЧУК
Луцький національний технічний університет

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ КОРЕЛЯЦІЙНОГО ТА РЕГРЕСІЙНОГО АНАЛІЗУ ТЕЛЕКОМУНІКАЦІЙНИХ МЕРЕЖ

Розглядаються методи дослідження взаємозв'язку між двома або більш неперервними змінними – регресійно-кореляційний аналіз з множинною та покроковою регресією. Запропоновано системний підхід до знаходження оптимального рішення, заснований на збалансованому виборі числа змінних і комплектних елементів вибірки. Розглянуті способи отримання статистичних оцінок параметрів мережі, основаних на методах видалення елементів, підстановки середнього, попарного викреслювання та підстановки регресії.

Ключові слова: *мультисервісна мережа, регресійно-кореляційний аналіз, множинна і покрокова регресія, правило зупинки.*

N.M. YAKYMCHUK
Lutsk National Technical University

COMPARATIVE ANALYSIS OF METHODS OF CROSS-CORRELATION AND REGRESSIVE ANALYSIS OF TELECOMMUNICATION NETWORKS

Two methods of studying the relationship between two or more continuous variables are considered - a regression-correlation analysis with multiple and stepwise regression. In a regression analysis a connection between a single variable called a dependent variable and several other called independent variables is considered. A systematic approach to finding an optimal solution based on a balanced choice of the following rule is proposed: the sample elements and/or variables with missing values should be deleted. So as to ensure a balance between the rest of the number of variables and the number of remaining elements, that is, to maximize the number of complete sample elements. There are considered the ways of receipt of statistical estimations of network parameters, based on the methods of delete of elements, substitution of middle, double deletion and substitution of regression/

Keywords*: multiservice network, regression-correlation analysis, multiple and stepwise regression, stop rule.*

Вступ. Мультисервісну мережу іноді називають мережею наступного покоління (англ. Next Generation Network, NGN). Концепція мереж NGN наступна. Змінюються самі принципи побудови систем зв'язку. Замість класичного поділу мереж зв'язку на первинну мережу і вторинні мережі сучасна архітектура мереж NGN включає в себе чотири рівні: послуг, управління, транспорту та доступу [1, 2].

Разом з тим, важливою проблемою під час проектування телекомунікаційних мереж є розробка алгоритмів, математичних методів і рівнянь, придатних для застосування під час розв'язання конкретних практичних мережних задач [3, 4].

Моделювання мережі математичними залежностями з детермінованими параметрами, тобто представлення мережі у вигляді детермінованої системи, в багатьох випадках є практично нездійсненним або дає занадто грубий, практично даремний результат з таких причин:

- отримання повної інформації про параметри і стан мережі в кожний момент часу в переважній більшості випадків практично неможливе;
- зміни стану і порушення в роботі мережі, перевантаження та інші аномальні ситуації є випадковими подіями, якими неможливо управляти – їх можна тільки прогнозувати з певною точністю.
- системи рівнянь для опису мережі будуть мати порядок, який можна порівняти з числом мережних і термінальних вузлів. Їх рішення в реальному часі потребує практично нереального обсягу обчислювальних ресурсів [5, 6], а помилки розрахунків будуть неприпустимо великі.

Тому в даний час тільки статистичні методи опису мереж, процесів обміну даними, синтезу структури мережі і оцінки параметрів, управління мережами можуть давати результати задовільної точності.

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

Метод множинної регресії. 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, $j = 1, 2, \ldots, N$.

Використання статистичних даних дозволяє домагатися оптимальних результатів, керуючи величинами факторів, або прогнозувати можливу величину результату при сформованих значеннях факторів. Загальне призначення множинної регресії полягає в аналізі зв'язку між кількома незалежними змінними (званими також регресорами) і залежної змінної [6].

Оцінювання проводиться за спостереженнями за входом (рядки матриці спостережень $\mathbf{X}$) і виходом (елементи вектора відгуків $\bar{\mathbf{y}}$).

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. Функціональна залежність результату від факторів представляється рівнянням регресії

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

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. 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 при і її подальшою перевіркою, необхідно прагнути, щоб рівняння включало якнайменше незалежних змінних.

Введемо поняття відсутніх значень. При використанні одновимірних за своєю природою методів аналізу (наприклад, t -критерію) найбільш розумний спосіб дії полягає у видаленні з вибірки елементів з відсутнім значенням $\boldsymbol{X}$ (аналізованої змінної). Однак ситуація змінюється при використанні істотно багатовимірних методів аналізу, тобто коли для кожного елемента вибірки є p спостережуваних змінних $X_1, X_2, \ldots, X_p$. Тепер, якщо елемент вибірки має відсутнє значення, скажімо, для змінної X_1, видалення цього елемента вибірки з аналізу не є необхідним, оскільки воно призводить до втрати інформації про змінні, що доставляється цим елементом. Так як множинний лінійний регресійний аналіз, так само як й інші багатовимірні процедури [5], засновані на векторі середніх *m* і матриці коваріацій *S*, можна залишити цей елемент у вибірці і використовувати наявні в ньому виміри для обчислення оцінок вектора середніх $\bar{\mathbf{x}}$ і матриці коваріацій **S**.

Оцінювання статистичних параметрів. Розглянемо тепер різні методи оцінювання $\bar{\mathbf{x}}$ та **S** (що еквівалентно, матриці кореляцій **R**), коли відсутні деякі значення [6].

Метод 1. Для обчислення оцінок $\bar{\mathbf{x}}$ та **S** використовуються тільки n_c комплектних елементів. Цей метод називається методом *видалення елементів*.

Метод 2. Для отримання використовуються спостереження. Замість відсутніх значень змінної підставляється її середня величина. Потім, використовуючи укомплектовану таким чином вибірку обсягу, отримують $\bar{\mathbf{x}}$ та **S**. Цей метод називається методом *підстановки середнього*.

Метод 3. Використовується n_i спостережень для отримання $\bar{x}_i$ та s_i й n_{ij} спостережень – для обчислення s_{ij}. Ці статистики служать компонентами $\bar{\mathbf{x}}$ та **S**.

Метод 4. Використовується n_i спостережень для отримання $\bar{x}_i$ та s_i й n_{ij} спостережень – для обчислення r_{ij}. Потім обчислюється значення s_{ij} як $s_{ij} = r_{ij} \times s_i \times s_j$, в чому і полягає відмінність даного методу від попереднього. Методи 3 і 4 називаються методами *попарного викреслювання*.

Метод 5. Використовується n_c комплектних елементів для оцінки регресії будь-якої змінної по всім іншим змінним. Наприклад, нехай рівняння регресії має вигляд $X_1 = f(X_2, \ldots, X_p)$. Тепер, якщо в *j*-му випадку є відсутнє значення X_1, воно замінюється оцінкою $\hat{x}_{1j} = f(x_{2j}, \ldots, x_{pj})$. Аналогічні рівняння можна отримати і для $X_2, \ldots, X_p$. Потім укомплектовані таким чином спостереження використовуються для обчислення $\bar{\mathbf{x}}$ та **S**.

Метод 6. На відміну від методу 5 для передбачення значення, наприклад, використовується або

одна змінна з $X_2,\ldots,X_p$, що найбільш корелювали з X_1, або деяка підмножина змінних з $X_2,\ldots,X_p$. Методи 5 і 6 носять назви методів *підстановки регресії*.

Основний недолік будь-якого з перерахованих методів пов'язаний з тим, що їх статистичні властивості за рідкісним винятком невідомі. Крім того, застосування таких методів часто призводить до зміщених оцінок.

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).

З огляду на все це можна дати наступну рекомендацію досліднику: елементи вибірки та/або змінні з відсутніми значеннями повинні бути видалені так, щоб забезпечити баланс між рештою числа змінних і числом елементів, що залишилися, тобто, максимізувати число комплектних елементів вибірки.

Отже, якщо елемент містить багато пропусків, його потрібно видалити. З іншого боку, слід видалити змінну, якщо її значення невідомо для більшості елементів. Після цього можна звичайним чином використовувати метод найменших квадратів або процедури багатовимірного статистичного аналізу.

Якщо число незалежних змінних велике, такий підхід для визначення найкращої підмножини практично не потрібен навіть при застосуванні ЕОМ. Наприклад, якщо p = 5, є всього 5 + 10 + 10 + 5 + 1 = 31 рівняння регресії, а якщо p = 10, то їх число становить вже 2× (10 + 45 + 120 + 210) + 252 + 1 = 1023. взагалі, коли число змінних дорівнює p, є $2p$–1 регресійних рівнянь. Обмеження на машинний час і допустимі витрати призводять до необхідності пошуку інших підходів.

Одним з рішень є покрокова регресія (пряма), коли незалежні змінні одна за одною включаються в підмножину згідно попередньо заданому критерію. У той же час деяка змінна може бути замінена іншою змінною, яка не входить в набір, або видалена з нього. Сукупність критеріїв, що визначають, які змінні включати, замінювати і видаляти, називається покроковою процедурою.

За допомогою покрокової процедури виходить упорядкований список предикторів. Наприклад, якщо p=5, такий список може мати вигляд X_2, X_5, X_1, X_4 і X_3. Для визначення "найкращої" підмножини з цього списку вибираються $m \le p$ перших змінних так, щоб

a) вони можливо краще передбачали $\mathbf{Y}$ і

b) їх число t було якомога менше.

Іншими словами, економний набір складається зі змінних впорядкованого списку, які мають найбільш високу здатність до прогнозування. У прикладі, наведеному вище, такий набір міг би складатися тільки з змінних X_2 і X_3, якби регресія по ним була майже такою ж "якісною", як і регресія з X_2, X_5, X_1, X_4 та X_3.

Процедура визначення числа t називається правилом зупинки. Таким чином, суть проведеного дослідження полягає саме в реалізації системного підходу. Методика безперервної діагностики мережі полягає в розбитті процесу на наступні взаємопов'язані етапи.

1. На першому етапі проводиться діагностика на фізичному рівні для виключення помилок і правильної інтерпретації результатів подальшого тестування.

2. На другому етапі доцільно проводити діагностику термінальних вузлів мережі шляхом стресового тестування мережі в двох режимах:

– режим калібрування з навантаженням тільки на мережу для виявлення помилок апаратної і програмної реалізації;

– режим з навантаженням тільки на мережу для виявлення проблем взаємодії станцій, вузьких місць на сервері і в каналах зв'язку.

3. На наступному етапі проводиться діагностика каналів зв'язку і серверів з використанням аналізаторів протоколів і аналізаторів серверів. Спільна обробка і аналіз отриманих в процесі тестування швидкісних характеристик, трендів характеристик мережного трафіку і лічильників серверів також здійснюється статистичними методами, що дозволяє встановити причини неправильного функціонування того чи іншого каналу зв'язку (сервера).

4. Заключний етап наскрізної діагностики мережі – діагностика прикладного мережного програмного забезпечення.

Висновки. Технологія моніторингу та аналізу телекомунікаційної мережі як великої та складної системи являє собою набір діагностичних засобів і методику їх використання, які дозволяють дати об'єктивну оцінку якості роботи прикладних програм в мережі і обґрунтувати рекомендації щодо поліпшення їх роботи. Концепція наскрізної діагностики мережі передбачає вміння ефективно оцінити, як працюють всі компоненти мережі з урахуванням їх взаємозв'язків і взаємовпливу. При цьому значна частина проблем функціонування мережі криється зовсім не у вичерпанні ліміту пропускної здатності, а в проблемах взаємодії апаратури, конфігурації, організації мережі і роботи користувачів.

Сьогодні все частіше системні адміністратори стикаються з проблемами в роботі додатків, викликані нераціональним використанням пропускної спроможності локальної мережі. Різного роду паразитний трафік здатний повністю поглинути все ресурси або поставити певні сервіси в невигідні умови роботи. У таких випадках доводиться вдаватися до рішень, що дозволяє розподіляти пропускні канали так,

щоб забезпечити роботу важливих для виробничого процесу додатків.

Важливим є також те, що схема управління трафіком на основі пріоритетів ще не повністю стандартизована, і, як правило, не підтримується додатками, виконуваними на окремих ПК. В таких умовах поставлену задачу легше і простіше перекласти на мережне комунікаційне обладнання.

Література

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 / Н. Н. Лесная // 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. – М. : 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.

References

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. Vinogradov N. A. Analysis of potential descriptions of commutation devices of and management of the new generations networks / N. A. Vinogradov // Zviazok. – 2004. – № 4. – P. 10-17.

4. Lesnaia N. N. Comparative analysis of estimation methods of intellectual network descriptions // Naukovi zapysky Ukrainskoho naukovo-doslidnoho instytutu zviazku. – 2009. – № 2(10). – P. 97-102.

5. Afifi A. Statisticalanalysis: using computer / A. Afifi, S. Eizen. – *Moskva: Mir,* 1982. 488 p.

6. Mosteller F. Data analysis and regression: issue 1 / F. Mosteller, J. Tyuki. – Moskva: Finansy i statistika, 1982. – 317 p.

7. Toroshanko Ya. I. Monitoring and analyse of telecommunication network parameters / Ya. I. Toroshanko, A. O. Bulakovska, M. S. Vysochinenko, V. S. Shmatko // Telekomunikatsiini ta informatsiini tekhnolohii. – 2014. – No. 3. – P. 62-69.

8. Barabash Yu. L. Questions of statistical recognition theory / Yu. L. Barabash, B. V. Varskyi, V. T. Zinov`ev, V. S. Kirichenko, V. F. Sapegin. – Moskva: Sovietskoe radio, 1967. – 400 p.

Рецензія/Peer review : 21.09.2017 р. Надрукована/Printed :31.10.2017 р.
Рецензент: д.т.н. Бойко Ю. М.