

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ОЦІНЮВАННЯ РОЗМІРУ JAVA ТА PHP-ЗАСТОСУНКІВ З ВІДКРИТИМ КОДОМ ЗА РЕГРЕСІЙНИМИ МОДЕЛЯМИ

Ворона М.В.¹, Приходько С.Б. д.т.н., професор², Смикодуб Т.Г.³
Національний університет кораблебудування імені адмірала Макарова

Україна, Миколаїв

¹ mvl.vorona@gmail.com, ² sergiy.prykhodko@nuos.edu, ³ tgsmyk@gmail.com

Анотація. Удосконалено трьохфакторну та чотирьохфакторну моделі нелінійної регресії для оцінювання розміру PHP та JAVA-застосунків з відкритим кодом на основі чотиривимірного та п'ятивимірного нормалізуючих перетворень Джонсона сім'ї S_B , що дозволяє підвищити достовірність оцінювання залежної змінної регресії у порівнянні з використанням одновимірних нормалізуючих перетворень.

Ключові слова: математичне моделювання; регресійна модель; оцінювання розміру PHP та JAVA-застосунків з відкритим кодом.

Вступна частина. Оцінювання розміру JAVA та PHP-застосунків з відкритим кодом як і іншого програмного забезпечення (ПЗ) на ранній стадії розробки є важливою задачею, оскільки ця інформація використовується для прогнозування трудомісткості створення ПЗ за допомогою такої відомої моделі як СОСОМО II. Це потребує відповідних математичних моделей для оцінювання розміру ПЗ, у тому числі і JAVA та PHP-застосунків з відкритим кодом.

Для оцінювання кількості строк коду інформаційних JAVA та PHP-систем з відкритим кодом відомі лінійні регресійні рівняння в залежності від трьох метрик концептуальної моделі даних у вигляді діаграми класів [1]. Ці рівняння побудовані на основі методів множинного лінійного регресійного аналізу. Але, як відомо, при побудові лінійних регресійних моделей необхідно виконання певних умов, зокрема залишки повинні бути розподілені за нормальним законом, що має місце лише в поодиноких випадках. А це веде до необхідності побудови нелінійних регресійних моделей для оцінювання кількості строк ПЗ із використанням відповідних методів математичного моделювання та множинного нелінійного регресійного аналізу [2].

Тому в [2, 3] для оцінювання розміру інформаційних JAVA та PHP-систем з відкритим кодом були запропоновані нелінійні регресійні моделі, які побудовані за допомогою множинного нелінійного регресійного аналізу із застосуванням чотиривимірного перетворення Джонсона сім'ї S_B на основі таких же трьох метрик діаграми класів, що і в [1]: загальна кількість класів, загальна кількість зв'язків та середня кількість атрибутів на клас. Але для JAVA та PHP-застосунків з відкритим кодом, що не є інформаційними системами, наприклад, таких як різноманітні фреймворки та конвертори, регресійні моделі можуть залежати в тому числі від інших метрик.

Ціллю роботи є удосконалення трьохфакторної та чотирьохфакторної моделі нелінійної регресії для оцінювання розміру PHP та JAVA-застосунків з відкритим кодом на основі чотиривимірного та п'ятивимірного нормалізуючих перетворень Джонсона сім'ї S_B , що дозволить підвищити достовірність оцінювання залежної змінної регресії у порівнянні з використанням одновимірних нормалізуючих перетворень.

Основна частина. В роботі для оцінювання розміру JAVA та PHP-застосунків з відкритим кодом побудовано дві багатофакторні нелінійні регресійні моделі. Для оцінювання розміру PHP-застосунків у тисячах строк коду побудовано трьохфакторну нелінійну регресійну модель в залежності від кількості класів (Classes) X_1 ; суми середньої кількості класів, на які впливає даний клас (Average Afferent Coupling) X_2 і середньої кількості класів, з яких даний клас отримує ефекти (Average Efferent Coupling), та середньої кількості методів (Average Methods) X_3 на основі чотиривимірного перетворення Джонсона сім'ї S_B . Для оцінювання розміру JAVA-застосунків побудовано чотирьохфакторну нелінійну регресійну модель в залежності від кількості класів X_1 ; кількості статичних методів (NOSM) X_2 ; метрики, що характеризує відсутність згуртованості методів (Lack of Cohesion of Methods, LCOM), X_3 та кількості викликів унікального методу в класі (the Response for Class, RFC) X_4 на основі п'ятивимірного перетворення Джонсона сім'ї S_B . При побудові моделей були використані дані із 38 Java-застосунків та 40 PHP-застосунків розташованих на сайті GitHub (<https://github.com>). Вибір саме цих факторів був обумовлений тим, що вони не мають проблеми з мультиколінеарністю: елементи на головній діагоналі оберненої коваріаційної матриці для зазначених відповідних факторів були менші за 6.

Нелінійна регресійна модель для оцінювання розміру PHP-застосунків має вигляд [2]

$$Y = \hat{\phi}_Y + \hat{\lambda}_Y \left[1 + e^{-(\hat{Z}_Y + \varepsilon - \hat{\gamma}_Y) / \hat{\eta}_Y} \right]^{-1}, \quad (1)$$

де ε – гаусівська випадкова величина, яка визначає залишки, $\varepsilon \sim N(0,1)$;

$$\hat{Z}_Y = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3, \quad \hat{b}_0 = 0, \quad \hat{b}_1 = 1,05746, \quad \hat{b}_2 = -0,0428158, \quad \hat{b}_3 = 0,504146;$$

$$Z_j = \gamma_j + \eta_j \ln \frac{X_j - \phi_j}{\phi_j + \lambda_j - X_j}, \quad \phi_j < X_j < \phi_j + \lambda_j, \quad j=1,2,3; \quad \hat{\gamma}_Y = 3,22402, \quad \hat{\gamma}_1 = 3,0914,$$

$$\hat{\gamma}_2 = 0,741344, \quad \hat{\gamma}_3 = 18,3888, \quad \hat{\eta}_Y = 0,673845, \quad \hat{\eta}_1 = 0,652695, \quad \hat{\eta}_2 = 0,880739, \quad \hat{\eta}_3 = 2,08381,$$

$$\hat{\phi}_Y = 0,054249, \quad \hat{\phi}_1 = -0,023277, \quad \hat{\phi}_2 = 1,52687, \quad \hat{\phi}_3 = -1,25874, \quad \hat{\lambda}_Y = 1057,484, \quad \hat{\lambda}_1 = 8737,832,$$

$$\hat{\lambda}_2 = 11,7836, \quad \hat{\lambda}_3 = 52797,72 \text{ [4]}.$$

Нелінійна регресійна модель для оцінювання розміру JAVA-застосунків має вигляд (1) тільки з тією різницею, що до неї входить ще четвертий фактор X_4 , а оцінки параметрів мають інші значення.

Нелінійна регресійна модель для оцінювання розміру PHP-застосунків на основі чотиривимірного перетворення Джонсона сім'ї S_B (1) виявилася кращою за модель, що було також побудовано за нормалізуючим перетворенням у вигляді десяткового логарифму, за двома показниками: R^2 і MMRE. Значення R^2 і MMRE для моделі на основі чотиривимірного перетворення Джонсона складають 0,982 і 0,161 відповідно, що краще за ці показники для моделі на основі десяткового логарифму відповідно на 1,0% і 7,1%. Для моделі на основі чотиривимірного перетворення Джонсона значення PRED(0,25) дорівнює 0,75, що на 3,3% гірше за цей показник для моделі на основі десяткового логарифму. Але основна перевага моделі на основі чотиривимірного перетворення Джонсона у порівнянні з моделлю на основі десяткового логарифму полягає у менших шириних інтервалу передбачення нелінійної регресії розміру PHP-застосунків з відкритим кодом для більшої кількості даних. Модель на основі чотиривимірного перетворення Джонсона сімейства S_B у порівнянні з моделлю на основі десяткового логарифму має менші ширини інтервалу передбачення для 36 PHP-застосунків.

Подібні результати ми отримали і для чотирьохфакторної нелінійної регресійної моделі для оцінювання розміру JAVA-застосунків на основі п'ятивимірного перетворення Джонсона сім'ї S_B [5]. Тільки з тією різницею, що її відсоток прогнозованих результатів $PRED(0,25)$ виявився кращим за відповідний показник попередньої моделі. Для чотирьохфакторної нелінійної регресійної моделі для оцінювання розміру JAVA-застосунків на основі п'ятивимірного перетворення Джонсона сім'ї S_B значення R^2 , MMRE та $PRED(0,25)$ є відповідно такими: 0,947, 0,175 та 0,824. Ці значення вказують на добру якість побудованої нелінійної моделі. Крім того, було побудовано чотирьохфакторну лінійну регресійну модель, для якої були отримані відповідно такі значення R^2 , MMRE та $PRED(0,25)$: 0,957, 0,683 та 0,529. Значення MMRE більше за 0,25 та $PRED(0,25)$ менше за 0,75 вказують на погану якість побудованої лінійної регресійної моделі для оцінювання залежної випадкової змінної. Добрий результат ця лінійна модель дає лише для оцінювання вибіркового середнього.

Отримані результати можна пояснити розподілом багатовимірних даних, який суттєво відрізняється від нормального. Зокрема про це свідчить оцінка багатовимірного ексцесу β_2 , яка визначалася як і в [2]. Відомо, що для m -вимірного нормального розподілу $\beta_2 = m(m+2)$. У разі п'ятивимірних даних $\beta_2 = 35$. У нашому випадку для даних із 38 Java-застосунків оцінка β_2 дорівнює 100,8, що майже в 3 рази перевищує теоретичне значення. І навпаки, для нормалізованих даних, які отримані за п'ятивимірним перетворенням Джонсона сімейства S_B , оцінка β_2 дорівнює 37,45, що всього на 7% відрізняється від теоретичного значення.

Кращі показники оцінювання розміру РНР-застосунків з відкритим кодом за моделлю нелінійної регресії (1) на основі чотиривимірного нормалізуючого перетворення Джонсона сім'ї S_B також можна в першу чергу пояснити кращою нормалізацією. Так, якщо за критерієм на основі квадрата відстані Махаланобіса гіпотеза про нормальність багатовимірного закону розподілу нормалізованих за допомогою чотиривимірного нормалізуючого перетворення Джонсона сім'ї S_B даних для 40 застосунків приймається для рівня значущості 0,025, то у випадку застосування одновимірного перетворення та без нього – відкидається.

Висновки. Удосконалено трьохфакторну модель нелінійної регресії для оцінювання розміру РНР-застосунків з відкритим кодом в залежності від загальної кількості класів; суми середньої кількості класів, на які впливає даний клас, і середньої кількості класів, з яких даний клас отримує ефекти, та середньої кількості методів на клас на основі чотиривимірного нормалізуючого перетворення Джонсона сім'ї S_B . Удосконалено чотирьохфакторну модель нелінійної регресії для оцінювання розміру JAVA-застосунків з відкритим кодом в залежності від загальної кількості класів; кількості статичних методів; метрики, що характеризує відсутність згуртованості методів, та кількості викликів унікального методу в класі на основі п'ятивимірного перетворення Джонсона сім'ї S_B . Це дозволяє підвищити достовірність оцінювання залежної змінної нелінійної регресії у порівнянні з використанням одновимірних нормалізуючих перетворень. Моделі, що побудовано на основі багатовимірного нормалізуючого перетворення Джонсона сім'ї S_B , в порівнянні з іншими регресійними моделями мають більший відсоток прогнозованих результатів, менші середні величини відносної похибки та ширини інтервалу передбачення нелінійної регресії.

REFERENCES

- [1] Tan, H.B.K., Zhao, Y., & Zhang, H. (2009). Conceptual data model-based software size estimation for information systems. *Transactions on Software Engineering and Methodology*, 2(19), October 2009, Article No. 4. DOI: 10.1145/1571629.1571630

- [2] Prykhodko, N.V., & Prykhodko, S.B. (2018). Constructing the non-linear regression models on the basis of multivariate normalizing transformations. *Electronic modeling*, 6(40), 101-110. DOI: 10.15407/emodel.40.06.101
- [3] Prykhodko, N.V., & Prykhodko, S.B. (2018). The non-linear regression model to estimate the software size of open source Java-based systems. *Radio Electronics, Computer Science, Control*, 3(46), 158-166. DOI: 10.15588/1607-3274-2018-3-17
- [4] Приходько, С.Б., Приходько, Н.В., Фаріонова, Т.А., & Ворона, М.В. (2020). Трьохфакторна нелінійна регресійна модель для оцінювання розміру Php-застосунків з відкритим кодом. *Науковий журнал «Вчені записки Таврійського національного університету імені В. І. Вернадського. Серія: Технічні науки»*. Том 31 (70) № 1, 124-131. DOI: 10.32838/2663-5941/2020.1-1/23
- [5] Приходько, С.Б., Приходько, Н.В., & Смикодуб, Т.Г. (2020). Чотирьохфакторна нелінійна регресійна модель для оцінювання розміру Java-застосунків з відкритим кодом. *Науковий журнал «Вчені записки Таврійського національного університету імені В. І. Вернадського. Серія: Технічні науки»*, Том 31 (70) № 2, 157-162. DOI: 10.32838/2663-5941/2020.2-1/25

Vorona M.V., Prykhodko S.B., Smykodub T.G.

Mathematical modeling of estimating the size of open source JAVA and PHP-applications using regression models

Abstract. Three-factor and four-factor nonlinear regression models for estimating the size of PHP and open source JAVA applications based on the four-variate and five-variate Johnson normalizing transformations for S_B family are developed, which improves the confidence of the dependent variable regression estimate compared to univariate normalizing transformations.

Keywords: mathematical modeling; regression model; estimation of the size of open source JAVA and PHP-applications.

Ворона М.В., Приходько С.Б., Смыкодуб Т.Г.

Математическое моделирование оценки размера JAVA и PHP-приложений с открытым кодом по регрессионным моделям

Аннотация. Усовершенствованы трехфакторная и четырехфакторная модели нелинейной регрессии для оценки размера PHP и JAVA-приложений с открытым кодом на основе четырехмерного и пятимерного нормализующих преобразований Джонсона семьи S_B , что позволяет повысить достоверность оценки зависимой переменной регрессии по сравнению с использованием одномерных нормализующих преобразований.

Ключевые слова: математическое моделирование; регрессионная модель; оценка размера JAVA и PHP-приложений с открытым кодом.