

ТРЬОХ-ФАКТОРНА РЕГРЕСІЙНА МОДЕЛЬ ДЛЯ ОЦІНЮВАННЯ РОЗМІРУ PHP-ЗАСТОСУНКІВ З ВІДКРИТИМ КОДОМ

Ворона М.В.¹, Приходько С.Б., д.т.н., професор²

^{1,2} Національний університет кораблебудування імені адмірала Макарова

^{1,2} Україна, Миколаїв

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

Анотація. Модель, довірчі інтервали та інтервали передбачення трьох-факторної нелінійної регресії для оцінювання розміру PHP-застосунків з відкритим кодом побудовані на основі чотиривимірного нормалізуючого перетворення Джонсона сім'ї S_B . Проведено порівняння побудованої моделі з відповідними нелінійними регресійними моделями отриманими на основі одновимірних нормалізуючих перетворень у формі десятикового логарифму і Джонсона сім'ї S_B .

Ключові слова: PHP, оцінювання розміру PHP-застосунку; нелінійна регресійна модель; довірчий інтервал; інтервал передбачення; нормалізуюче перетворення.

PHP – це популярна мова сценаріїв загального призначення з відкритим вихідним кодом, яка особливо добре підходить для розробки веб-застосунків на стороні сервера. В наш час PHP використовується більш ніж у 80% всіх веб-сайтів, наприклад, таких як Tesla, Wikipedia, WordPress.com [1]. Хоча основне призначення цієї мови полягає у тому, щоб дозволити веб-розробникам швидко писати веб-сторінки, що генеруються динамічно, але за допомогою PHP роблять набагато більше, в тому числі різноманітні фреймворки, конвертори, HTTP-клієнти, системи управління фотографіями і відправки електронної пошти, застосунки для обробки зображень і виявлення мобільних пристроїв та багато іншого.

Задача оцінювання розміру програмного забезпечення (ПЗ) на ранніх стадіях його проектування є важливою, оскільки результати цього оцінювання використовуються для прогнозування трудомісткості з розробки різноманітного ПЗ, включаючи PHP-застосунки з відкритим кодом.

У роботах [2, 3] запропоновано рівняння лінійної регресії для оцінювання розміру ПЗ інформаційних систем з відкритим кодом на Java та PHP. Запропоновані рівняння побудовано за допомогою множинного лінійного регресійного аналізу на основі метрик з діаграми класів. Проте існують чотири основні припущення, які виправдовують використання моделей лінійної регресії, одним з яких є нормальність розподілу помилок. Але це припущення справедливе лише в окремих випадках. Це призводить до необхідності використання нелінійних регресійних моделей, в тому числі для оцінювання розміру PHP-застосунків з відкритим кодом.

Для побудови нелінійної регресійної моделі для оцінювання розміру PHP-застосунків з відкритим кодом були використані дані з метрик 40 застосунків розташованих на сайті GitHub (<https://github.com>): кількість строк коду, кількість класів, кількість зв'язків, кількість методів та кількість доступів до атрибутів. Ці дані були отримані за допомогою двох інструментів PhpMetrics та PhpLoc. Зазначимо, що вказані дані мають негаусівський розподіл. Для їх нормалізації було використано чотиривимірне перетворення Джонсона сім'ї S_B .

В роботі для оцінювання розміру РНР-застосунків з відкритим кодом пропонується використовувати трьох-факторну нелінійну регресійну модель, що побудована на основі чотиривимірного перетворення Джонсона сім'ї S_B [4]:

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

де $\hat{Z}_Y$ – результат оцінювання за лінійним регресійним рівнянням $\hat{Z}_Y = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \hat{b}_3 Z_3$; Z_1 , Z_2 і Z_3 – нормалізовані змінні відповідно факторів X_1 , X_2 і X_3 , $Z_j = \gamma_j + \eta_j \ln[(X_j - \phi_j) / (\phi_j + \lambda_j - X_j)]$, $\phi_j < X_j < \phi_j + \lambda_j$, $j = 1, 2, 3$; $\hat{b}_0 = 0$; $\hat{b}_1 = 0,40264$; $\hat{b}_2 = 0,96674$; $\hat{b}_3 = 0,04678$; $\hat{\gamma}_Y$, $\hat{\gamma}_1$, $\hat{\gamma}_2$, $\hat{\gamma}_3$, $\hat{\eta}_Y$, $\hat{\eta}_1$, $\hat{\eta}_2$, $\hat{\eta}_3$, $\hat{\phi}_Y$, $\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\phi}_3$, $\hat{\lambda}_Y$, $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\lambda}_3$ – оцінки параметрів чотиривимірного перетворення Джонсона сім'ї S_B відповідно змінних Y , X_1 , X_2 і X_3 ; $\hat{\gamma}_Y = 1,93903$; $\hat{\gamma}_1 = 3,30679$; $\hat{\gamma}_2 = 2,17555$; $\hat{\gamma}_3 = 8,75135$; $\hat{\eta}_Y = 0,64128$; $\hat{\eta}_1 = 1,00076$; $\hat{\eta}_2 = 0,62267$; $\hat{\eta}_3 = 1,79318$; $\hat{\phi}_Y = -0,10319$; $\hat{\phi}_1 = -1,11934$; $\hat{\phi}_2 = -0,87768$; $\hat{\phi}_3 = -0,17518$; $\hat{\lambda}_Y = 210,889$; $\hat{\lambda}_1 = 367,259$; $\hat{\lambda}_2 = 2851,784$; $\hat{\lambda}_3 = 215,249$; ε – випадкова величина з розподілом Гаусу, $\varepsilon \sim N(0,1)$.

Нелінійна регресійна модель (1) дозволяє оцінювати розмір РНР-застосунків з відкритим кодом в тисячах рядків коду Y в залежності від трьох факторів: середня кількість доступів до атрибутів на клас X_1 , загальна кількість класів X_2 та середня кількості зв'язків на клас X_3 .

Значення відсотка прогнозування PRED (0,25) для моделі (1) краще (0,600) в порівнянні з відповідними значеннями для нелінійних моделей побудованих на основі одновимірних нормалізуючих перетворень у формі десяткової логарифму (0,525) і Джонсона сім'ї S_B (0,550). Крім того, для моделі (1) ширини інтервалів передбачення та довірчих інтервалів виявилися меншими для більшої кількості даних.

Висновки.

Удосконалено трьох-факторну нелінійну регресійну модель для оцінювання розміру РНР-застосунків з відкритим кодом в залежності від метрик діаграми класів на основі чотиривимірного перетворення Джонсона сім'ї S_B , що дозволяє підвищити достовірність оцінювання залежної змінної нелінійної регресії у порівнянні з лінійною регресійною моделлю та нелінійними моделями, що побудовані на основі одновимірних нормалізуючих перетворень у формі десяткової логарифму і Джонсона сім'ї S_B .

REFERENCES

- [1] Hayden, J. (2019) 80% of the web powered by PHP [Electronic resource] / James Hayden. Access mode: <https://haydenjames.io/80-percent-web-powered-by-php/>
- [2] Tan, H.B.K., Zhao, Y., & Zhang, H. (2006). *Estimating LOC for information systems from their conceptual data models*. Proceedings of the 28th International Conference on Software Engineering (ICSE '06), Shanghai, China, May 20-28, 2006.
- Kaczmarek, J., & Kucharski, M. (2004). Size and effort estimation for applications written in Java. *Information and Software Technology*, 46, (9), 589-601. DOI: 10.1016/j.infsof.2003.11.001
- [3] Kiewkanya, M., & Surak, S. (2016). *Constructing C++ software size estimation model from class diagram*. Computer Science and Software Engineering: 13th International Joint Conference, Khon Kaen, Thailand, July 13-15, 2016. DOI: 10.1109/JCSSE.2016.7748880

[4] Prykhodko, N.V., & Prykhodko, S.B. (2018). Constructing the non-linear regression models on the basis of multivariate normalizing transformations. *Electronic modeling*, 40, (6), 101–110. DOI: 10.15407/emodel.40.06.101

Vorona M.V., Prykhodko S.B.

A three-factor regression model to estimate the size of open-source PHP-based apps

Annotation. *The model, confidence and prediction intervals of three-factor non-linear regression for estimating the size of open-source PHP-based apps are constructed on the basis of the Johnson four-dimensional normalizing transformation for S_B family. Comparison of the constructed model with the linear regression model and non-linear regression models based on the univariate normalizing transformations, both the decimal logarithm and the Johnson transformation for S_B family is performed.*

Keywords: PHP, PHP-based application size estimation; non-linear regression model; confidence interval; prediction interval; normalizing transformation.

Ворона М.В., Приходько С.Б.

Трехфакторная регрессионная модель для оценки размера РНР-приложений с открытым исходным кодом

Аннотация. *Модель, доверительные интервалы и интервалы прогнозирования нелинейных регрессий для оценки размера РНР-приложений с открытым кодом построены на основе нормализующего преобразования четырехмерного нормализующего преобразования Джонсона семейства S_B . Проведено сравнение построенных моделей с соответствующими линейной регрессионной моделью и нелинейными регрессионными моделями на основе одномерных нормализующих преобразований в виде десятичного логарифма и преобразования Джонсона семейства S_B .*

Ключевые слова: РНР, оценка размера РНР-приложения, нелинейная регрессионная модель, доверительный интервал, интервал прогнозирования, нормализующее преобразование.