

ПЕРЕВІРКА ЯКОСТІ РЕГРЕСІЙНОЇ МОДЕЛІ ОЦІНЮВАННЯ ТРУДОМІСТКОСТІ РОЗРОБКИ ВЕБ-СКРАПЕРІВ НА JAVA

Цуман Д.С.¹, Латанська Л.О., к.ф.-м.н., доцент², Каіров В.О., к.т.н.³
^{1,2,3} Національний університет кораблебудування імені адмірала Макарова

^{1,2,3} Україна, м. Миколаїв

¹ pilgrim.tsu@gmail.com, ² llatanskaya@gmail.com, ³ vladimir.kairov@gmail.com

Анотація. Побудована нелінійна регресійна модель для оцінювання трудомісткості розробки веб-скраперів на Java. Виконана перевірка якості розробленої моделі шляхом обчислення коефіцієнту детермінації, суми квадратів відхилень, середньої величини відносної похибки, рівня прогнозування.

Ключові слова: трудомісткість; нелінійна регресійна модель; якість регресійної моделі; веб-скрапери.

Вступна частина. Прогнози щодо точності оцінювання трудомісткості розробки програмного забезпечення існуючими методами та моделями зазвичай правдиві лише для певного типу проектів, аналогічних тим, на яких ті чи інші моделі чи методи були випробувані. Тобто, не існує універсальних методів чи моделей, які являються оптимальними для всіх проектів і середовищ розробки [1].

Тому є необхідність удосконалювати існуючі та розробляти нові методи та моделі оцінювання трудомісткості створення проектів обраних типів, зокрема веб-скраперів на Java. Для розроблених чи вдосконалених методів та моделей обов'язковою є перевірка їх якості. Це дозволяє підвищити точність оцінювання трудомісткості розробки відповідних проектів, що робить задачу актуальною [2].

Метою роботи є підвищення точності оцінювання трудомісткості розробки веб-скраперів на Java.

Основна частина. Нелінійна регресійна модель оцінювання трудомісткості розробки програмного забезпечення побудована на основі наступних факторів:

- кількість рядків коду (X);
- трудомісткість розробки веб-скраперів на Java (Y).

Досліджено, що розподіл X, Y не є гаусівським. Тому для побудови нелінійної регресійної моделі виконали нормалізацію емпіричних даних (нормалізуюче перетворення – десятковий логарифм) [3].

Лінійна регресійна модель для нормалізованих даних Z_x, Z_y має вигляд:

$$\hat{Z}_y = b_0 + b_1 \cdot Z_x + \varepsilon, \quad (1)$$

де b_0, b_1 – коефіцієнти лінійної регресії, які обчислюються методом найменших квадратів;

ε – випадкова помилка.

Шляхом використання зворотнього перетворення отримали нелінійну регресійну модель оцінювання трудомісткості розробки веб-скраперів на Java:

$$\hat{Y} = e^{b_0 + \varepsilon} \cdot X^{b_1}. \quad (2)$$

Для перевірки якості розробленої регресійної моделі використовувались наступні показники: коефіцієнт детермінації R^2 , сума квадратів відхилень S_y , середня величину відносної похибки $MMRE$, рівень прогнозування $PRED(l)$ [4].

- Коефіцієнт детермінації R^2 характеризує частку дисперсії результативної ознаки, що пояснюється регресією, в загальній дисперсії результативної ознаки та визначається за формулою:

$$R^2 = 1 - (\sum_1^n (y_i - \hat{y}_i)^2 / \sum_1^n (y_i - \bar{y})^2), \quad (3)$$

де y_i – емпіричне значення;

$\hat{y}_i$ – розраховане значення;

$\bar{y}$ – середнє значення.

Коефіцієнт детермінації приймає значення від 0 (відсутній лінійний зв'язок між показниками) до 1 (відсутній кореляційний зв'язок між показниками). При значенні $R^2 \geq 0,5$ можна вважати, що дана модель є прийнятною, при $R^2 \geq 0,8$ – достатньо ефективною.

- Сума квадратів відхилень вибірових значень від розрахованих S_y використовується для перевірки якості рівнянь регресії та порівняння різних моделей:

$$S_y = \sum_1^n (y_i - \hat{y}_i)^2. \quad (4)$$

- Середня величина відносної похибки $MMRE$ визначається за формулою:

$$MMRE = \frac{1}{n} \cdot \sum_1^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \quad (5)$$

Прийнятне значенням $MMRE \leq 0,25$.

- Рівень прогнозування $PRED(l)$ обчислюється за формулою:

$$PRED(l) = \frac{k}{n}, \quad (6)$$

де k – кількість значень $\left| \frac{y_i - \hat{y}_i}{y_i} \right| \leq l$ ($i = \overline{1, n}$).

Зазвичай $l = 0,25$. Задовільне значенням $PRED(0,25) \geq 0,75$.

Побудована модель має прийнятні значення обчислених величин та може бути використана для прогнозування.

Висновки. В результаті проведеного дослідження були побудовані лінійна (1) для нормалізованих даних та нелінійна (2) через зворотне перетворення регресійні моделі оцінювання трудомісткості розробки веб-скраперів на Java. Перевірка якості та адекватності розробленої моделі здійснювалася шляхом обчислення коефіцієнту детермінації (3), суми квадратів відхилень (4), середньої величини відносної похибки (5), рівня прогнозування (6). В подальших дослідженнях планується розробити програмне забезпечення для реалізації моделі.

REFERENCES

[1] Титов, А. И. (2016). Выбор метрики размера проекта в модели оценки трудоемкости разработки программ. *Интеллектуальные технологии на транспорте*, 1, 31-38.

- [2] Malathi, S. et al (2012). Analysis of size metrics and effort performance criterion in software cost estimation. *Indian Journal of Computer Science and Engineering*, 3 (1), 24-31.
- [3] Демиденко, Е.З. (1981). *Линейная и нелинейная регрессии*. М.: Финансы и статистика.
- [4] Гмурман, В.Е. (2010). *Теория вероятностей и математическая статистика. Учеб. пособие*. М.: Высшее образование.

Dmytro Tsuman, Liudmyla Latanska, Kairov Volodymyr

Checking the quality of the regression model for estimating the labor intensity of developing web scrapers in Java

Abstract. A nonlinear regression model is built to estimate the labor intensity of developing web scrapers in Java. The quality of the developed model was checked by calculating the coefficient of determination, the sum of squares of deviations, the mean value of the relative error, the level of prediction.

Keywords: labor intensity; nonlinear regression model; quality of the regression model; web scrapers.

Цуман Д.С., Латанская Л.А., Каиров В.А.

Проверка качества регрессионной модели оценивания трудоемкости разработки веб-скраперов на Java

Аннотация. Построена нелинейная регрессионная модель для оценивания трудоемкости разработки веб-скраперов на Java. Выполнена проверка качества разработанной модели путем вычисления коэффициента детерминации, суммы квадратов отклонений, средней величины относительной погрешности, уровня прогнозирования.

Ключевые слова: трудоемкость; нелинейная регрессионная модель; качество регрессионной модели; веб-скраперы.