УДК 004.412:519.237.5

ВИЯВЛЕННЯ ВИКИДІВ У ДВОВИМІРНИХ ДАНИХ МЕТРИК CBO ТА RFC ЗАСТОСУНКІВ З ВІДКРИТИМ КОДОМ НА KOTLIN ЗА ДОПОМОГОЮ МАТЕМАТИЧНИХ МОДЕЛЕЙ

Приходько С.Б., д.т.н., професор[1]; Кольцов А.В.[2]
Національний університет кораблебудування імені адмірала Макарова
Україна, Миколаїв
[1] sergiy.prykhodko@nuos.edu.ua; [2] andrew.koltsv@gmail.com

Анотація. *У роботі запропоновано математичні моделі для підвищення достовірності визначення викидів у двовимірних даних із метрик CBO та RFC програмних застосунків мовою Kotlin. Аналіз 102 проєктів з відкритим кодом на Kotlin показав суттєві відмінності від Java-застосунків, що потребувало створення окремої моделі. Для врахування негаусівського розподілу метрик використано двовимірне нормалізуюче перетворення Бокса-Кокса, після чого побудовано моделі виявлення викидів на основі χ^2- та F-розподілів. Запропоновані моделі підвищують достовірність виявлення викидів у порівнянні з існуючими моделями для мови Java.*
Ключові слова: *метрики програмного забезпечення; CBO; RFC; викиди; Kotlin; перетворення Бокса-Кокса; відстань Махаланобіса; еліпс прогнозування.*

Вступна частина. Визначення складності та якості програмного забезпечення (ПЗ) є однією із важливих задач програмної інженерії, в тому числі в умовах поширення ПЗ з відкритим кодом. Для цього широко застосовують метрики програмного забезпечення, зокрема набір метрик CK (Chidamber & Kemerer metrics suite) [1]. Серед них CBO (Coupling Between Objects) та RFC (Response For Class) характеризують взаємозв'язки між класами й дозволяють кількісно оцінювати складність об'єктно орієнтованого проєктування. У реальних наборах даних метрики часто мають розподіл, що відхиляється від нормального. Таке відхилення ускладнює виявлення викидів і знижує достовірність моделей, побудованих на класичних припущеннях про нормальність. Це зумовлює потребу у створенні нових моделей, що враховують негаусівську природу емпіричних даних.

Метою роботи є вдосконалення математичних моделей для визначення викидів у двовимірних даних із метрик CBO та RFC програмних застосунків з відкритим кодом на Kotlin з урахуванням відхилення розподілу від нормального на основі двовимірного нормалізуючого перетворення Бокса-Кокса з урахуванням тестової статистики на основі квантиля розподілу хі-квадрат і тестової статистики на основі квантиля F-розподілу Фішера.

Основна частина. У роботі використано емпіричні дані 102 проєктів з відкритим кодом на Kotlin, для яких за допомогою інструменту CodeMR (https://www.codemr.co.uk/) отримано значення метрик CBO та RFC. Порівняльний аналіз із даними для Java-застосунків показав значну різницю як у середніх значеннях, так і в характері розподілу метрик.

Результати тесту Мардія [2] підтвердили істотну негаусівськість розподілу: значення двовимірної асиметрії $\beta_1 = 70{,}867$ та ексцесу $\beta_2 = 14{,}066$ перевищують критичні значення (9,30 та 9,48 відповідно).

Для подолання цього обмеження застосовано двовимірне нормалізуюче перетворення Бокса–Кокса, що дозволило отримати перетворені дані, розподіл яких не відхилявся від нормального. За нормалізованими даними побудовано дві моделі для визначення викидів у

вигляді еліпсів прогнозування із застосуванням тестових статистик на основі відповідно квантилів χ^2-розподілу та F-розподілу Фішера з рівнем значимості 0,005.

Обидві моделі базуються на використанні методів на основі квадрату відстані Махаланобіса для нормалізованих даних [3, 4], що враховує кореляцію між метриками та дозволяє виявляти віддалені спостереження в двовимірному просторі та враховувати відхилення розподілу від нормального. Розрахунки показали відсутність викидів після застосування запропонованих моделей. При цьому модель із тестовою статистикою на основі χ^2-розподілу виявилася більш суворою, що виражено через еліпс меншої площі.

У результаті було усунуто недолік моделі для Java [5] щодо спроби її використання для застосунків на Kotlin, де значна частина Kotlin-проєктів (близько 25%) хибно ідентифікувалася як викиди через відмінність розподілу метрик. Розроблені моделі зберігають геометричну інтерпретацію через еліпс прогнозування для нормалізованих метрик, що є важливим для візуального аналізу даних.

Висновки. У роботі побудовано вдосконалені математичні моделі для визначення викидів у двовимірних даних із метрик CBO та RFC програмних застосунків на мові Kotlin. Моделі базуються на поєднанні нормалізуючого перетворення Бокса–Кокса та статистик для χ^2- і F-розподілів, що дозволяє враховувати негаусівський характер даних. Отримані результати демонструють вищу достовірність порівняно з моделями для Java та підтверджують доцільність створення наведених моделей.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Chidamber, S. R., & Kemerer, C. F. (1994). A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, 20(6), 476–493. https://doi.org/10.1109/32.295895.

[2] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530. https://doi.org/10.1093/biomet/57.3.519.

[3] Prykhodko, S., Prykhodko, N., Makarova, L., & Pukhalevych, A. (2018). Application of the squared mahalanobis distance for detecting outliers in multivariate non-Gaussian Data. *2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)*, 962–965. https://doi.org/10.1109/tcset.2018.8336353.

[4] Prykhodko, S., Prykhodko, N., Makarova, L., Kudin, O., Smykodub, T., & Prykhodko, A. (2017). Detecting bivariate outliers on the basis of normalizing transformations for non-Gaussian data. *Proceedings of the Vth International Conference "Advanced Information Systems and Technologies"*, Sumy, Ukraine, 2017, pp. 95–97.

[5] Prykhodko, S., Prykhodko, N., & Smykodub, T. (2022). A joint statistical estimation of the RFC and CBO metrics for open-source applications developed in Java. *2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*, 442–445. https://doi.org/10.1109/csit56902.2022.10000457.

Prykhodko Sergiy, Koltsov Andrii

Detection of outliers in two-dimensional CBO and RFC metrics data of open-source applications in Kotlin using mathematical models

Abstract. *The study proposes mathematical models aimed at improving the confidence of outlier detection in two-dimensional data derived from the CBO and RFC metrics of Kotlin-based software applications. An analysis of 102 open-source projects in Kotlin revealed significant differences compared to Java applications, which necessitated the development of a dedicated model. To correct the non-Gaussian distribution of metrics, the Box–Cox normalization transformation was applied, followed by the construction of outlier detection models based on χ^2 and F distributions. The proposed models enhance the confidence of outlier identification compared to existing Java-oriented models.*

Keywords: *software metrics; CBO; RFC; outliers; Kotlin; Box-Cox transformation; Mahalanobis distance; prediction ellipse.*