УДК 004.412

МАТЕМАТИЧНА МОДЕЛЬ ДЛЯ ВИЗНАЧЕННЯ СУПРОВОДЖУВАНОСТІ MACHINE LEARNING JAVA-ЗАСТОСУНКІВ

Трофименко О.Г.[1]; Фаріонова Т.А., к.т.н., доц.[2]
[1,2]Національний університет кораблебудування імені адмірала Макарова
[1,2]Україна, Миколаїв
[1]ORCID: 0000-0001-7626-0886; [2]ORCID 0000-0003-3384-4712

Анотація. *Побудовано математичну модель для визначення супроводжуваності machine learning Java-застосунків у вигляді довірчих інтервалів та інтервалів передбачення нелінійних регресійних моделей для метрик RFC, CBO і WMC на основі перетворення Бокса-Кокса за даними 109 застосунків.*
Ключові слова: *супроводжуваність; machine learning Java-застосунки; метрики RFC, CBO, WMC; нелінійна регресійна модель; інтервал передбачення.*

Вступ. Java залишається однією з основних мов корпоративного розроблення, і її дедалі частіше використовують для реалізації функціональності машинного навчання. За даними звіту Azul State of Java 2026, 62 % респондентів застосовують Java для розроблення функціональності штучного інтелекту, а 31 % організацій повідомили, що таку функціональність містить понад половина їхніх Java-застосунків [1]. На відміну від традиційних програмних систем, machine learning Java-застосунки змінюються не лише через нові вимоги, а й через оновлення даних, перенавчання моделей і появу нових версій бібліотек. Sculley та співавтори [2] пов'язують цю особливість із накопиченням прихованого технічного боргу: код самої моделі становить невелику частину системи, а основний обсяг припадає на інфраструктуру збирання даних, виділення ознак і конфігурування, що ускладнюється через інтеграційний код (glue code) і заплутані конвеєри обробки даних. Дослідження інженерів Microsoft [3] також показало, що компоненти систем машинного навчання значно складніше ізолювати один від одного, ніж компоненти традиційного програмного забезпечення. Тому для machine learning Java-застосунків важливо завчасно визначати супроводжуваність, тобто виявляти можливі труднощі внесення змін ще до того, як вони потребуватимуть значних ресурсів.

Особливості об'єктно-орієнтованого коду, що впливають на супроводжуваність, характеризують метрики набору Чідамбера–Кемерера [4]: зв'язаність класів CBO (Coupling Between Object classes), множина відгуку класу RFC (Response For a Class) і зважена кількість методів класу WMC (Weighted Methods per Class). Проте окремі значення цих метрик і фіксовані порогові значення не враховують взаємозв'язків між метриками: наприклад, зростання складності класів зазвичай супроводжується зростанням їхньої множини відгуку. Цей недолік усунуто в методиці С. Приходька [5], за якою значення кожної метрики порівнюється з довірчим інтервалом та інтервалом передбачення нелінійної регресійної моделі цієї метрики від двох інших. Однак модель у [5] побудована за даними 51 Java-застосунку загального призначення, а архітектурні особливості machine learning Java-застосунків можуть зумовлювати інші розподіли метрик і закономірності їхнього взаємозв'язку. Тому використання загальної моделі може призводити до некоректних висновків, і для machine learning Java-застосунків потрібна окрема модель.

Мета дослідження – підвищення достовірності визначення супроводжуваності machine learning Java-застосунків за рахунок побудови математичної моделі у вигляді довірчих інтервалів та інтервалів передбачення нелінійних регресійних моделей для метрик RFC, CBO і WMC на основі тривимірного нормалізуючого перетворення Бокса-Кокса.

Основна частина. Для побудови моделі сформовано вибірку зі 109 machine learning Java-застосунків з відкритим кодом: 66 бібліотек машинного навчання, 33 прикладні застосунки та 10 фреймворків великих мовних моделей. Для кожного застосунку інструментом статичного аналізу CK обчислено середні за класами значення RFC, CBO і WMC. За тестом Мардіа розподіл вихідних даних не є тривимірним нормальним: на рівні значущості 0,05 статистика асиметрії перевищує квантиль розподілу хі-квадрат, а статистика ексцесу – квантиль стандартного нормального розподілу. Тому дані нормалізовано тривимірним перетворенням Бокса-Кокса з параметрами, оціненими методом максимальної правдоподібності: $\lambda_{\text{RFC}} = 0{,}1644$, $\lambda_{\text{CBO}} = 0{,}6948$, $\lambda_{\text{WMC}} = 0{,}0920$.

Типові значення метрик machine learning Java-застосунків (RFC ≈ 11,7; CBO ≈ 5,7; WMC ≈ 12,4) помітно менші, ніж у Java-застосунків загального призначення (16,5; 11,5; 16,9), за даними яких побудовано модель у роботі [5]. Через це частину machine learning Java-застосунків ця модель визначає як багатовимірні викиди: для 17 застосунків статистика, обчислена за квадратом відстані Махаланобіса, перевищила квантиль F-розподілу $F_{3;\,48;\,0{,}005} = 4{,}85$ [5], тобто модель [5] до них не застосовна. Серед фреймворків великих мовних моделей таких застосунків 5 із 10.

Після нормалізації багатовимірних викидів не виявлено: максимальне значення статистики 3,96 менше за квантиль $F_{3;\,106;\,0{,}005} = 4{,}53$. Нормалізовані дані задовольняють обидві умови тесту Мардіа: статистики асиметрії та ексцесу не перевищують відповідних квантилів. За нормалізованими даними методом найменших квадратів побудовано три лінійні регресійні моделі, з яких після оберненого перетворення Бокса-Кокса отримано нелінійні регресійні моделі для RFC, CBO і WMC:

$$Y_{RFC} = [0{,}164(-6{,}263 + 0{,}228\,\mathrm{X}_{CBO}^{0{,}695} + 6{,}77\,\mathrm{X}_{\mathrm{WMC}}^{0{,}092} + \varepsilon 1) + 1]^{1/0{,}164},$$

$$Y_{CBO} = [0{,}695(1{,}137 + 6{,}88\,\mathrm{X}_{RFC}^{0{,}164} - 6{,}403\,\mathrm{X}_{WMC}^{0{,}92} + \varepsilon 2) + 1]^{1/0{,}695},$$

$$Y_{WMC} = [0{,}092(-6{,}543 + 6{,}729\,\mathrm{X}_{RFC}^{0{,}164} - 0{,}211\mathrm{X}_{CBO}^{0{,}695} + \varepsilon 3) + 1]^{1/0{,}092},$$

де ε_1, ε_2, ε_3 – гаусівські випадкові величини з нульовим математичним сподіванням. Показники якості регресійних моделей (коефіцієнт детермінації R^2, середня величина відносної похибки MMRE та частка прогнозів з відносною похибкою не більше 0,25 PRED(0,25)) наведено в таблиці 1.

Таблиця 1 – Показники якості нелінійних регресійних моделей (N = 109)

Залежна змінна	Незалежні змінні	R^2	MMRE	PRED(0,25)
RFC	CBO, WMC	0,724	0,146	0,807
CBO	RFC, WMC	0,190	0,195	0,771
WMC	RFC, CBO	0,693	0,245	0,642

Для кожної регресійної моделі обчислено межі довірчого інтервалу та інтервалу передбачення з рівнем значущості 0,05. Рівень супроводжуваності визначається так: якщо значення всіх трьох метрик лежать у довірчих інтервалах, він середній; якщо всі вони лежать між нижніми межами інтервалів передбачення і верхніми межами довірчих інтервалів, але не всі в довірчих інтервалах, – високий; інакше – низький. За побудованою моделлю високий рівень супроводжуваності визначено для 12 застосунків вибірки, середній – для 2, низький – для 95. Для 84 із 95 застосунків з низьким рівнем усі метрики лежать в інтервалах передбачення, тобто низький рівень зумовлений не аномальністю застосунку, а тим, що одна з метрик перевищує типове значення за двома іншими. Саме ця метрика вказує, що слід покращувати: множину відгуку класів, їхню зв’язаність чи складність методів.

Висновки. Визначення супроводжуваності machine learning Java-застосунків є актуальним завданням, оскільки такі застосунки накопичують технічний борг через залежність від даних, великий обсяг інтеграційного коду та взаємозалежність компонентів. За методикою [5] на основі тривимірного перетворення Бокса-Кокса за даними 109 застосунків побудовано математичну модель для визначення супроводжуваності machine learning Java-застосунків у вигляді довірчих інтервалів та інтервалів передбачення нелінійних регресійних моделей для метрик RFC, CBO і WMC. Регресійні моделі для RFC і CBO задовольняють критеріям MMRE ≤ 0,25 і PRED(0,25) ≥ 0,75, регресійна модель для WMC – тільки критерію MMRE. Подальші дослідження передбачають покращення регресійної моделі для WMC, перевірку моделі на незалежній вибірці та розроблення програмного забезпечення для автоматизації розрахунків відповідних метрик і визначення рівня супроводжуваності за вихідним кодом застосунку.

Використання інструментів ШІ. Інструмент ШІ Gemini використовувався авторами для пошуку джерел та стилістичного редагування тексту. Усі наукові результати, модель і висновки сформульовані авторами самостійно.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Deslandes, N. (2026). *What Azul's 2026 State of Java report tells us about enterprise development.* TechInformed. https://techinformed.com/what-azuls-2026-state-of-java-report-tells-us-about-enterprise-development/

[2] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28* (pp. 2503–2511). Curran Associates.

[3] Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (pp. 291–300). IEEE. https://doi.org/10.1109/ICSE-SEIP.2019.00042

[4] 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

[5] Prykhodko, S. (2024). Evaluating quality of software systems by the confidence and prediction intervals of regressions for RFC, CBO and WMC metrics. *WSEAS Transactions on Systems, 23*, 322–330. https://doi.org/10.37394/23202.2024.23.36

Trofymenko O.G., Farionova T.A.

Mathematical Model for Determining the Maintainability of Machine Learning Java Applications

Abstract. *A mathematical model for determining the maintainability of machine learning Java applications has been developed in the form of confidence and prediction intervals of nonlinear regression models for the RFC, CBO, and WMC metrics based on the Box–Cox transformation using data from 109 applications.*

Keywords: *maintainability; machine learning Java applications; RFC, CBO, WMC metrics; nonlinear regression model; prediction interval.*