УДК 004.415.2.045 (076.5)

О.П. Дишлевий, асп.

ПРЕДМЕТНО-ОРІЄНТОВАНИЙ МЕТОД ПОБУДОВИ ЗАЛЕЖНОСТЕЙ МІЖ МЕТРИКАМИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Запропоновано схему предметно-орієнтованого методу побудови залежностей між метриками програмного забезпечення. Розглянуто середовища для побудови залежностей. Зроблено висновок про потребу в розробці спеціального середовища для побудови залежностей між метриками програмного забезпечення. Описано проведені експерименти. Визначено особливості методу.

The article is dedicated to designing of a subject-oriented method of making dependences between software metrics. Existing environments for making dependences are reviewed, and the conclusion about need for development of special environment for making dependences between software metrics is made. Article describes the experiments which defined specifics of a method, offers its scheme, and defined special features of the method.

емпірична інженерія програмного забезпечення, кореляційно-регресійний аналіз, метрика

Постановка проблеми

Інженерія програмного забезпечення розглядає питання досліджень програмного забезпечення, які належать до емпіричної інженерії програмного забезпечення [1]. Основним методом досліджень є вимірювання. Один з головних інструментів цих досліджень – метрики. За допомогою метрик оцінюють властивості складових розробки програмного забезпечення (продуктів, процесів).

Існують прямі та непрямі метрики [2]. Прямі метрики піддаються вимірюванню, але їх недостатньо для оцінки більшості властивостей. Тоді використовуються непрямі метрики, які формуються на основі прямих. Головне завдання – визначати вид та ступінь залежності непрямої метрики від прямої [3]. Для цього використовують два підходи – статистичний аналіз [4] та нейронні мережі [5].

Для визначення залежностей між метриками за допомогою нейромереж слід не тільки знати значення метрик (вхідних величин), а й провести навчання нейромережі [5]. Навчання нейромережі проводиться на уже отриманих раніше даних. Сформулювати правила для навчання нейромережі неможливо, оскільки визначенням залежностей між метриками досі не займалися. Для визначення залежностей між метриками за допомогою статистичного аналізу достатньо знати тільки значення метрик, а залежності будуються статистичними методами [4], тому необхідно застосовувати статистичний аналіз.

Статистичний аналіз проводиться за допомогою відповідних математичних програмних середовищ, до яких належать MatLab, MathCad, Maple, Mathematica, MS Excel. Крім них можна використати статистичні програмні середовища загального призначення Statistica, SPSS, SAS, Systat, Minitab, Statgraphics чи програмні середовища спеціального призначення SYSTAT, S-plus, STATA, PRISM, STADIA, Олимп, Класс-Мастер, Статистик-Консультант.

Досвід їх використання свідчить, що математичні програмні середовища та статистичні програмні середовища загального призначення для вирішення поставленої задачі потребують додаткового програмування з використанням статистичних алгоритмів. Середовища для емпіричних досліджень в програмному забезпеченні немає.

Отже, для визначення залежностей між метриками програмного забезпечення потрібне середовище та метод, який буде використовуватися в середовищі.

У статті пропонується метод обробки даних емпіричних досліджень програмного забезпечення за допомогою статистичного аналізу та засіб – статистичне середовище спеціального призначення, яке реалізує метод.

Розроблення методу

Серед методів емпіричних досліджень програмного забезпечення визначають [3]:

- керовані експерименти;
- дослідження ситуацій;
- дослідження-огляди;
- етнографії;
- дослідження дій.

Задача визначення залежностей між метриками програмного забезпечення належить до дослідження-огляду [3]. У рамках дослідження відбувається вимірювання прямих метрик програмного забезпечення, збір даних-результатів вимірювань, оброблення даних та визначення залежностей непрямих метрик від прямих. Розроблення методу визначення залежностей між метриками програмного забезпечення здійснюється шляхом статистичного аналізу з урахуванням особливостей програмного забезпечення.

Статистичний аналіз, який виконується з метою визначення залежностей, складається з трьох етапів [4]:

− первинний статистичний аналіз;

− кореляційний аналіз;

− регресійний аналіз.

Залежності будуються на етапах кореляційного та регресійного аналізу, але зробити це без попереднього аналізу даних на першому етапі неможливо, як це застосовується в різних прикладних науках [6–8]. Це пов'язано з тим, що метою першого етапу є визначення виду розподілу досліджуваної величини, від якого залежать наступні етапи. Існує велика кількість розподілів [9], але для визначення подальших досліджень суттєву роль відіграє наявність чи відсутність нормального розподілу [10]. Залежно від наявності нормального розподілу використовуються ті чи інші алгоритми побудови залежностей, тому важливо, чи мають нормальний розподіл досліджувані метрики.

Дослідження залежностей у прикладних науках показало, що для кожної із них існують свої закономірності, притаманні тільки даним з їх предметної області [6; 7; 8], тому був проведений експеримент з визначення закону розподілу метрик програмного забезпечення в рамках дослідження повторного використання програмного забезпечення [11].

Метою дослідження було визначення програмного коду, придатного для повторного використання. Висновок про можливість повторного використання можна зробити тоді, коли програмний код буде відповідати певним критеріям. За критерій був взятий набір метрик: контролю, метрики даних та типографські метрики [2; 12].

Метрики контролю – це метрики, які відображають складність модуля.

Метрики даних – це метрики, які відображають кількість та процентне співвідношення змінних, викликів підпрограм та кількість їх параметрів.

Типографські метрики – це метрики, які відображають коментованість програмного коду та осмисленість в назвах програмних конструкцій.

Всього було досліджено 52 метрики.

Задачею експерименту з визначення законів розподілу було розрахувати оптимальні значення для розглянутих метрик та можливий діапазон їх відхилень. Для цього виміряли значення усіх метрик. Для визначення оптимальної кількості значень метрик під час побудови законів розподілу вимірювання проводилося в декілька етапів.

Спочатку виміряли п'ятсот програм. Для кожної метрики розрахували математичні характеристики та їх відхилення [13]. Далі виміряли ще сто програм. Для значень метрик, отриманих з усіх програм, знову провели аналогічні розрахунки. Таким чином виміряли тисячу програм та провели подібні обчислення. Виходячи із закону великих чисел [9], зроблено висновок про недоцільність подальшого збільшення програм для вимірювань, так як статистичні характеристики мають практично однакові значення.

На наступному етапі побудовано для кожної метрики закон розподілу та обчислено відхилення. Якщо значення метрики для програми потрапляло в допустимий інтервал відхилень, то робили висновок про можливість повторного використання виміряного програмного коду. Коли значення метрики виходило за межі інтервалу, то робили висновок про неможливість повторного використання виміряного програмного коду.

Отримані типові закони розподілу метрик показано на рис. 1.

Аналіз всіх гістограм показує, що їх вигляд можна звести до таких чотирьох типових:

− гістограми з лівою асиметрією (рис. 1, *а*, *б*);

– унімодальна гістограма з сильною лівою асиметрією (рис. 1, *в*);

– багатомодальна гістограма (рис. 1, *г*).

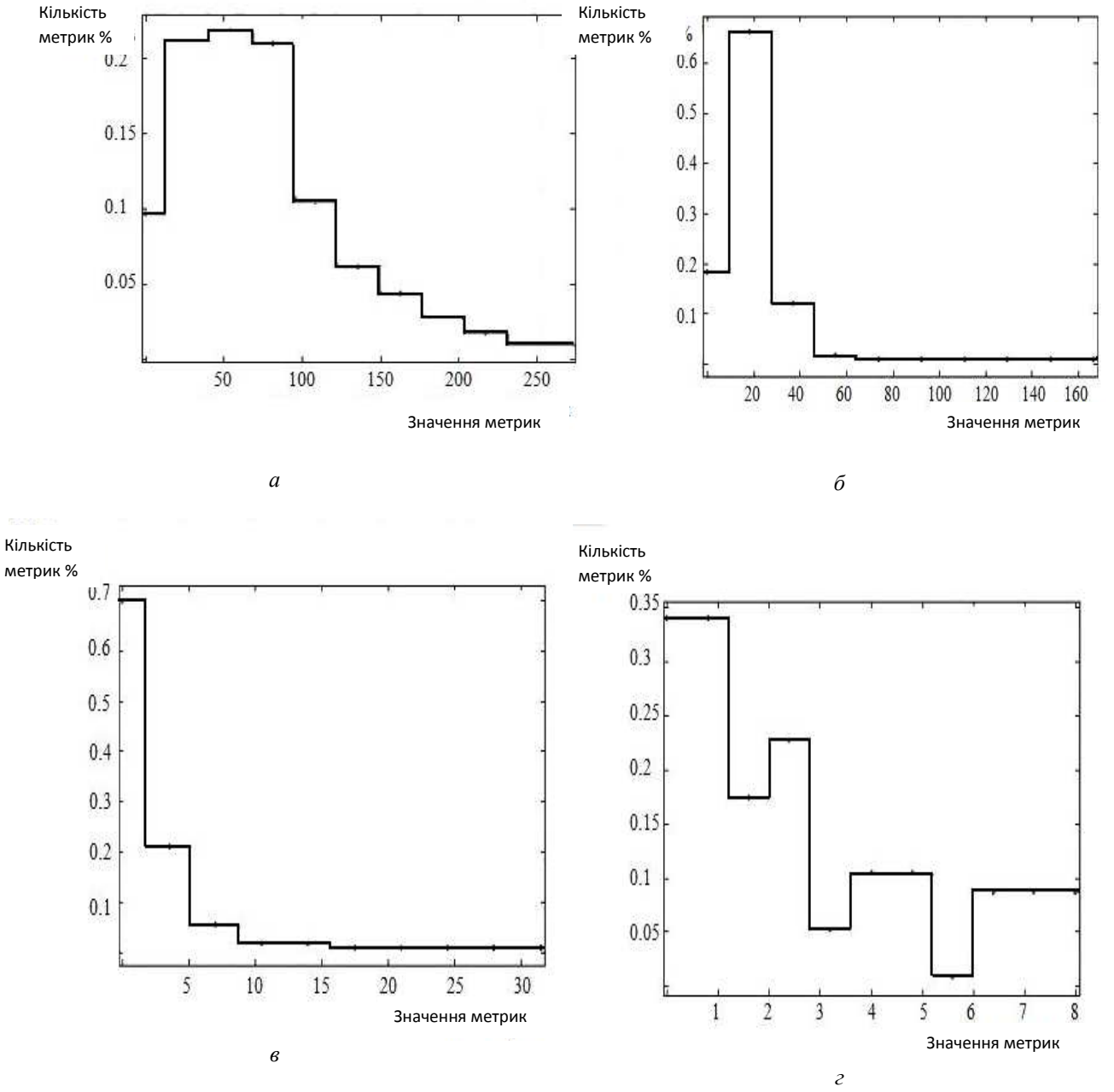


Рис. 1. Гістограми метрик:
а − максимальна кількість непустих рядків у модулі;
б − максимальна кількість викликів інших функцій, обчислене в модулі;
в − середня кількість викликів вводу – виводу, використаних в кожному модулі;
г − середня кількість аргументів (параметрів), використаних в кожному модулі

Експеримент показав дві особливості програмного забезпечення:

− доступність великої кількості програм для дослідження (проблема в інших науках через складність отримання даних), що дає можливість використовувати точні, а не наближені методи розрахунків;

− відсутність нормального закону розподілу метрик.

Результати експерименту показали: оскільки у метрик немає нормального закону розподілу, то визначати закон розподілу в розроблюваному методі недоцільно, але необхідно обов'язково перевіряти об'єм вибірки.

Для визначення наявності залежності непрямої метрики від прямої проводиться кореляційний аналіз двома шляхами [10]:

− простий розрахунок коефіцієнтів парної кореляції, коли досліджувані величини мають нормальний розподіл;
− розрахунок парної рангової кореляції, коли нормального закону розподілу немає.

Для емпіричних досліджень програмного забезпечення з визначення залежності непрямої метрики від прямої потрібно використати розрахунок парної рангової кореляції, що пов'язано з розподілом відмінним від нормального.

Відмінність парної рангової кореляції полягає в порівнянні не самих значень величин чи їх статистичних характеристик, а рангів, тобто номерів досліджуваних величин у відповідних матрицях (наборах метрик). Визначається парна рангова кореляція методом обчислення коефіцієнта Спірмена чи Кендала [14]. Залежність для значень коефіцієнтів відмінних від − 1, 0 та 1 підтверджується розрахунком значущості.

Під час проведення розрахунків не потрібно перевіряти точність отриманих значень, оскільки вимірювання метрик програмного забезпечення не мають похибок, пов'язаних з людським фактором чи засобом вимірювань.

Отже, кореляційний аналіз у розроблюваному методі повинен дозволяти:
− проводити розрахунок парної рангової кореляції;
− не перевіряти на точність отримані дані.

Для визначення виду залежності непрямої метрики від прямої застосовується регресійний аналіз. Він полягає у побудові та розрахунках коефіцієнтів функції регресії, яка відображає залежність непрямої метрики від прямої. Виділяють два види регресій [14]:
– лінійну;
− нелінійну.

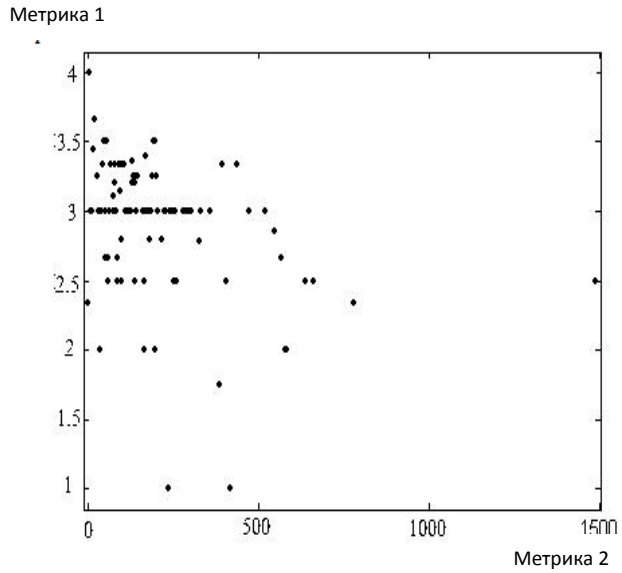
Лінійна регресія будується у випадку, коли при кореляційному аналізі було зроблено висновок про наявність лінійної залежності, інакше – нелінійна регресія. Лінія регресії будується на основі кореляційного поля. Кореляційні поля метрик програмного забезпечення зводяться до двох видів (рис. 2).

Якщо побудовані точки кореляційного поля потрапляють у коло (рис. 2, *а*), залежності немає.

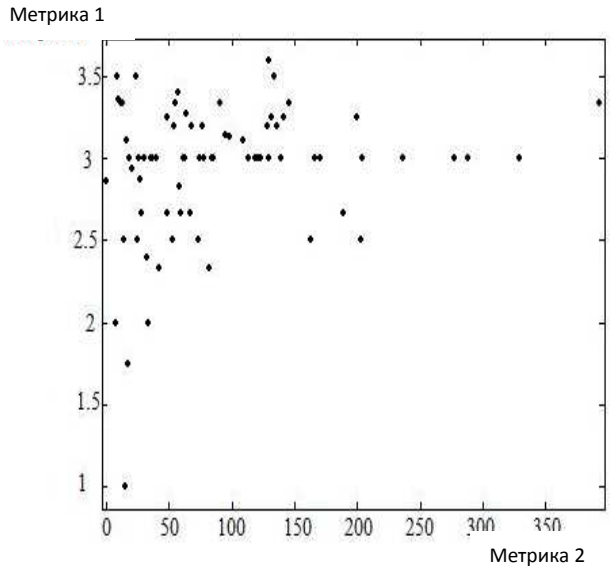
Якщо ж кореляційне поле не вписується у коло (рис. 2, *б*), а має інший вигляд, то робиться висновок про нелінійну залежність в лінії регресії.

Оскільки дані досліджень програмного забезпечення не мають нормального закону розподілу, то будується нелінійна регресія. Обов'язковою передумовою побудови будь-якої регресії є нормальний закон розподілу залежної метрики або обох метрик, якого немає.

У зв'язку з великою вибіркою ця передумова ігнорується. Єдиної теорії побудови нелінійної регресії немає [14], тому під час регресійного аналізу залежно від даних використовується той чи інший метод нелінійної регресії.



а



б

Рис. 2. Кореляційні поля:
а − вписується в коло;
б − складної конфігурації

Для реалізації регресійного аналізу в розроблюваному методі нелінійної регресії використали порівняння методів на даних досліджень програмного забезпечення. Будувалися наближені лінії регресії методом найменших квадратів, поліномів Чебишева та лінеаризації (побудова найпростіших наближених функцій).

Суть порівняння полягає в побудові ліній регресії різними способами для кожного з кореляційних полів з подальшим визначенням найточнішої лінії. У зв'язку з великим об'ємом проведення розрахунків прийнято рішення про побудову лінії регресії спочатку для однієї прямої метрики та непрямої метрики «простота сприйняття», а далі для контролю отриманих даних побудувати лінії регресії для декількох метрик. Кількість метрик збільшували доти, доки не підтвердили закономірність.

Спочатку для побудови лінії регресії була взята метрика «середнє значення непустих рядків у модулі». Її вибір пов'язаний з великим значенням значущості, що говорить про її суттєвий вплив на непряму метрику «простота сприйняття» [13]. Для неї побудовано наближені функції регресії трьома способами.

Виявилося, що максимальний степінь функції – 3, що говорить про простоту функції регресії. Її слід вибирати серед невеликого переліку простих функцій [14]. Після розрахунків коефіцієнтів наближених функцій та перевірки відхилення функцій виявилося, що найоптимальнішою функцією є експоненційна функція (рис. 3).

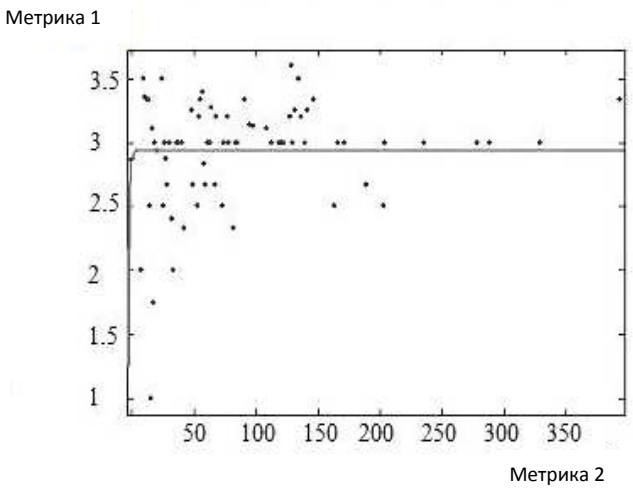


Рис. 3. Лінія регресії прямої метрики «середнє значення непустих рядків в модулі» та непрямої метрики «простота сприйняття»

Далі побудували лінії регресії ще для трьох метрик. Результати побудови та розрахунків коефіцієнтів лінії регресії для цих метрик підтвердили відсутність складних функцій залежності з великими степенями. Максимальний степінь залишився 3.

Результати порівняння методів нелінійної регресії показали, що перші два методи недоцільно використовувати для побудови лінії регресії, тому що максимальний степінь найближчої наближеної лінії регресії − 3. Тому для реалізації запропонованого методу було обрано метод лінеаризації.

На основі проведених досліджень пропонують метод побудови залежностей між метриками програмного забезпечення на основі статистичного аналізу.

Запропонований метод дозволяє визначати залежність непрямої метрики від прямої без попереднього визначення їх законів розподілу та будувати лінію регресії без попереднього аналізу досліджуваних величин.

Отже, сутність предметно-орієнтованого методу побудови залежностей між метриками програмного забезпечення полягає в тому, що побудова залежностей відбувається статистичними методами з врахуванням високої точності вимірювань програмного забезпечення без похибок та гіперболічною спадною залежністю між значеннями метрик та кількістю виміряних програм.

Таким чином, запропонований метод побудови залежностей між метриками програмного забезпечення має вигляд, показаний на рис. 4.

Його можна використовувати для метрик контролю, метрик даних та типографських метрик, дослідження яких було описано раніше. Для інших метрик можна стверджувати, що ситуація буде аналогічною. Але для підтвердження потрібно проводити додатковий первинний статистичний аналіз.

Висновки

В основі запропонованого предметно-орієнтованого методу побудови залежностей між метриками програмного забезпечення лежить використовуваний в інших галузях кореляційно-регресійний метод статистичних досліджень. Науковою новизною запропонованого методу є:

− виявлення й пояснення закону розподілу метрик відмінного від нормального;

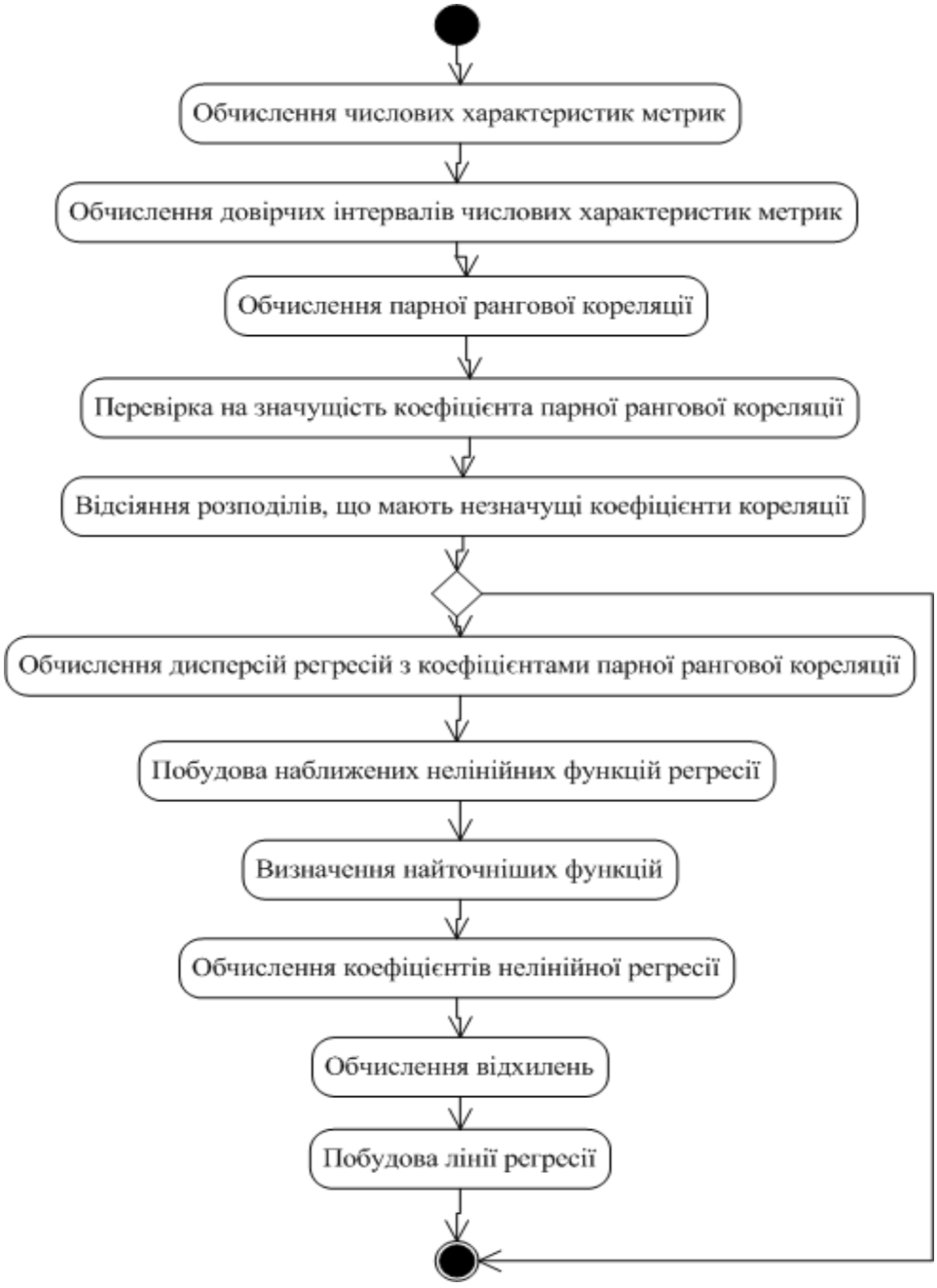


Рис. 4. Предметно-орієнтований метод побудови залежностей між метриками програмного забезпечення

− вилучення первинного статистичного аналізу з досліджень через відсутність нормальних розподілів серед досліджуваних метрик;
− умова про обов'язкове велике значення вибірки;
− використання в кореляційному аналізі метрик тільки парної рангової кореляції;
− відсутність перевірок точності коефіцієнтів кореляції через точність отриманих даних вимірювань;
− відсутність перевірки спільного закону розподілу метрик;
− наявність тільки нелінійної функціональної залежності між метриками простого виду (степінь функції не більший ніж 3);
− побудова регресії методом лінеаризації.

Розроблений метод допомагає досліднику програмного забезпечення зрозуміти суть та особливості цього дослідження.

Метод являє собою чітку послідовність дій, які повинен виконати дослідник програмного забезпечення під час визначення залежності непрямої метрики від прямої.

Література

1. Koji Torii. Ginger 2: An Enviroment for Computer-Aided Empirical Software Engineering / Torii Koji, Kenichi Matsumoto, Kumiyo Nakakoji at al/// IEEE Transactions on Software Engineering, Vol 25. No 4. July − August 1999, P 475−486.

2. Norman E. Fenton. Software Metrics: A Rigorous and Practical Approach / Norman E. Fenton, Shari Lawrence Pfleeger. − Cambridge University Press,1996. − 638p.

3. Forrest Shull. Guide to Advanced Empirical Software Engineering / Forrest Shull, Janice Singer, Dag I.K. Sjoberg. – Springer-Verlag London Limited 2008. − 394p.

4. Вентцель Е.С. Теория вероятностей: учеб. для вузов / Е.С. Вентцель. – 7-е изд. стер. – М.: Высш. шк., 2001. – 575 с.: ил.

5. Уоссермен Ф. Нейрокомпьютерная техника: Теория и практика. / Ф. Уоссермен / пер. с англ. – М.: Мир, 1992. − 118 с.

6. Рокицкий П.Ф. Биологическая статистика / П.Ф. Рокицкий. − Изд. 3-е, испр. – Минск: Высшэйш. школа, 1973. − 320 с.

7. Айвазян С. А. Прикладная статистика и основы эконометрики: учеб. для вузов / С.А. Айвазян, В.С. Мхитарян. − М.: ЮНИТИ, 1998. − 1022 с.

8. Дружинин Н.К. Математическая статистика в экономике / Н.К. Дружинин. − М.:Статистика, 1971. – 262 с.

9. Кендалл М. Теория распределений / М. Кендалл, А. Стюарт / пер. с англ. − М.: Наука, Глав. ред. физ.-мат. лит., 1966. – 588 с.

10. Кендалл М. Статистические выводы и святи / М. Кендалл, А. Стюарт. – М.: Наука, Глав. ред. физ.-мат. лит., 1973. – 899 с.

11. Хоменко В.А. Метод экспертного оценивания свойств повторно используемых компонентов программного обеспечения / В.А. Хоменко, А.П. Дышлевый // Матеріали VIII Міжнар. наук.-техн. конф. «Авіа-2007». − К.:НАУ, 2007, Т.1. − С. 13.169 − 13.172.

12. Изосимов А.В. Метрическая оценка качества программ / А.В. Изосимов, А.Л. Рыжков. – М.: МАИ, 1989. − 96 с.

13. Дишлевий О.П. Перевірка адекватності метричних моделей властивостей програмного забезпечення / А.П. Дишлевий // Інженерія програмного забезпечення 2006: Матеріали Всеукр. конф. аспірантів та студентів. − К.:НАУ, 2007. − С.77−84.

14. Айвазян С.А. Прикладная статистика: Исследование зависимостей: справ. изд. / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин / под. ред. С.А. Айвазяна. – М.:Финансы и статистика, 1985. – 487 с.

Стаття надійшла до редакції 07.09.09.