

ВИКОРИСТАННЯ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ ПРИ ПОБУДОВІ МОДЕЛЕЙ ТРИВАЛОСТІ ПРОЄКТІВ СТВОРЕННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Приходько С.Б.¹, д.т.н., проф.; Кудін О.О.¹, к.т.н.

¹Національний університет кораблебудування імені адмірала Макарова

Україна, Миколаїв

¹sergiy.prykhodko@nuos.edu.ua; ¹oleg.kudin@nuos.edu.ua

Анотація. В наш час розробники програмного забезпечення використовують математичні моделі часу виконання проєктів, основані на історичних даних. Але вони мають незадовільні показники якості. Для покращення показників якості означених моделей запропоновано перед побудовою моделі розділити загальний набір даних на кластери та побудувати моделі для кожного з них.

Ключові слова: регресійний аналіз; регресійні моделі часу проєктів; кластерний аналіз; метод K-Means; ітеративні методи кластеризації.

Вступна частина. В наш час при плануванні розкладу проєктів використовуються математичні моделі тривалості проєктів, які базуються на історичних даних. Наприклад, такі моделі, як СОСОМО, СОСОМО II або ISBSG. Практика показує, що використання історичних даних дозволяє побудувати означені моделі, але ці моделі мають низкі показники якості – середню величину відносної похибки MMRE та процент передбачених значень Pred(0,25). Одним зі способів покращення моделей і, відповідно, їх показників якості є розподіл загального набору даних на кластери та побудова моделі для кожного кластера [1].

Мета роботи. Метою дослідження є перевірка можливостей розбиття відомих наборів даних з програмних метрик на окремі кластери та побудова регресійних моделей для кожного з кластерів.

Основна частина. Для прогнозування тривалості нових проєктів програмного забезпечення (далі – проєктів ПЗ) сьогодні зазвичай використовують методи регресійного аналізу, які дозволяють побудувати регресійні моделі тривалості в залежності від трудомісткості на основі баз даних виконаних проєктів. При використанні означених баз даних слід мати на увазі, що різні проєкти мали різні значення трудомісткості та, відповідно, тривалості. Проєкти ПЗ створювались з використанням різних мов програмування та були призначені для вирішення задач відмінних за обсягами та складністю обчислень. Відповідно, побудовані регресійні моделі мали незадовільні показники якості. Підвищити показники якості регресійних моделей можна завдяки розподілу даних по групах відповідно до трудомісткості і часу виконання, а потім побудувати моделі для кожної групи окремо. Для вирішення цієї задачі є сенс використати методи кластерного аналізу.

В наш час для кластеризації даних часто використовують метод K-Means [1, 2]. Перевагою методу є проста реалізація. Але при його використанні потрібно заздалегідь визначити кількість кластерів, що в деяких випадках проблематично. Також були випадки коли кластери частково перетинались між собою. Тому існує актуальна проблема щодо визначення можливостей та обмежень використання методу K-Means при кластеризації бази даних проєктів ПЗ відповідно до трудомісткості і часу виконання.

Для вимірювання відстані між об'єктами в методі K-Means при кластеризації зразків з безперервними ознаками використовується квадратична евклідова відстань між двома точками x та y в m -вимірному просторі:

$$d(x, y)^2 = \sum_{j=1}^m (x_j - y_j)^2 = \|x - y\|_2^2,$$

де j відноситься до j -ого виміру (стовпчику ознаки) вхідних зразків x та y .

Якщо брати за основу метрику евклідової відстані, то алгоритм K-Means можна описати як просту задачу оптимізації – ітераційний підхід для мінімізації внутрикластерної суми квадратичних похибок (Sum of Squared Errors – SSE)

$$SSE = \sum_{i=1}^n \sum_{j=1}^k w^{(i,j)} \|x^{(i)} - \mu^{(j)}\|_2^2,$$

де $\mu^{(j)}$ - репрезентативна точка (центроїд) для кластера j ; $w^{(i,j)} = 1$, якщо зразок $x^{(i)}$ знаходиться в кластері j , та $w^{(i,j)} = 0$ в іншому випадку.

Для виявлення мінімального значення SSE потрібно виконати декілька ітерацій кластеризації набору даних починаючи з $k = 1 + p$ при $p = 0, 1, 2, \dots, p_{max}$. Припинення ітерацій ($k = p_{max}$) відбудеться, коли SSE перестане змінюватись.

Алгоритм методу має простий вигляд:

1. Перед обчисленням дослідник призначає кількість кластерів (k);
2. Випадково призначаються центроїди (центроїдами називають центральні точки відповідних кластерів) кожному з k кластерів;
3. Обчислюється відстань усіх спостережень до кожного з k центроїдів;
4. Призначаються спостереження до найближчого центроїда;
5. Знаходиться нове розташування центроїда, взявши середнє значення всіх спостережень у кожному кластері;
6. Повторюйте кроки 3 – 5, доки центроїди не змінять положення.

В дослідженні було використано відомий набір даних ISBSG6 [3] для персональних комп'ютерів (PC), який використовували багато авторів [3]. Він складається з даних 52 проєктів ПЗ. В наборі даних є такі параметри проєктів: час виконання проєктів (місяці); трудомісткість виконання проєктів (людино-години); тип комп'ютерів (MF - мейнфрейм, MR – комп'ютери середнього класу, PC - персональні комп'ютери).

За значенням SSE було визначено, що статистичну вибірку можна розділити на 2 кластери (рис. 1).

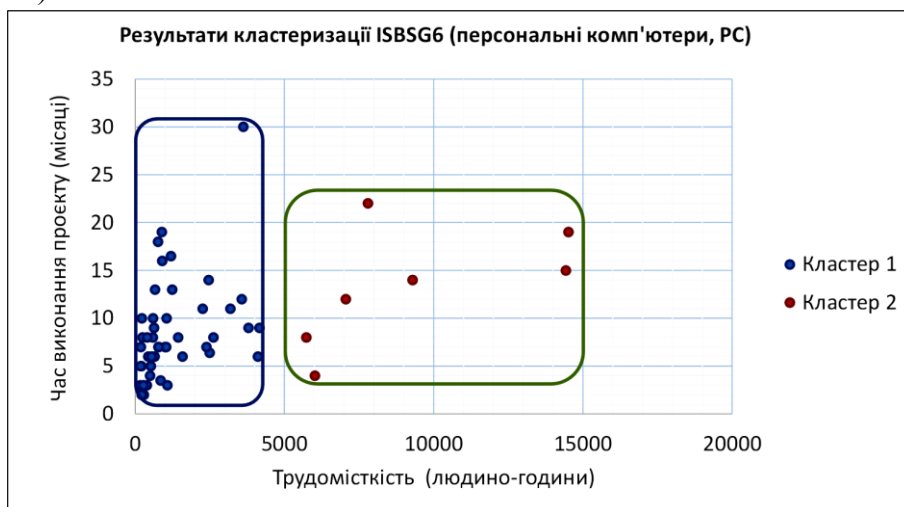


Рисунок 1 – Результати кластеризації статистичної вибірки ISBSG6 (PC)

На рис. 1 можна побачити, що вибірка розділена на 2 кластери, які не перетинаються між собою. Отже, існує дві групи проєктів ПЗ з різними значеннями трудомісткості робіт. Для цих груп можна побудувати дві окремі регресійні моделі.

При проведенні дослідження було побудовано дві регресійні моделі для кластера 1 (рис. 1) з використанням нормалізуючого перетворення на основі десяткового логарифму (рис. 2) та одновимірного перетворення Джонсона сімейства S_B (рис. 3).

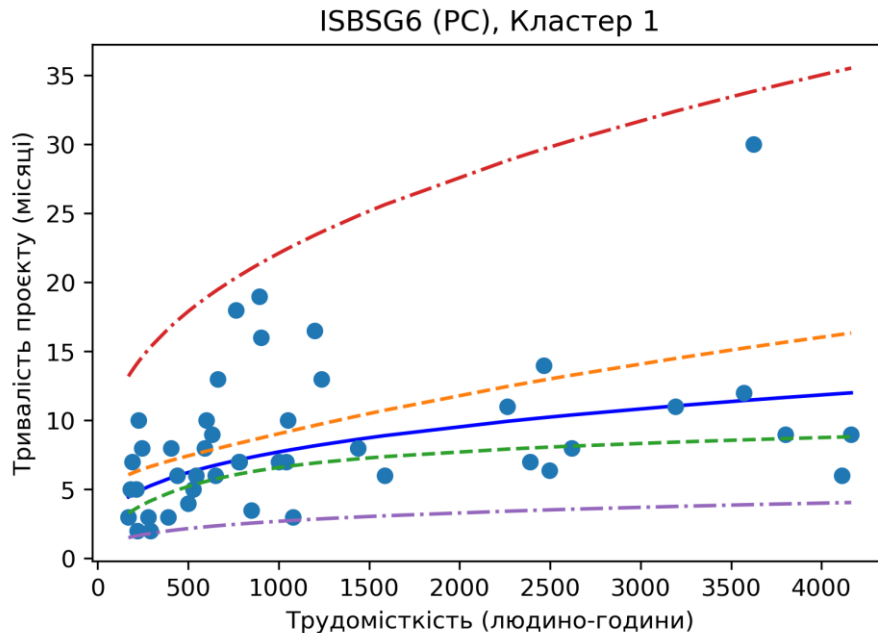


Рисунок 2 – Регресійна модель тривалості проєктів ПЗ, побудована з використанням нормалізуючого перетворення на основі десяткового логарифму

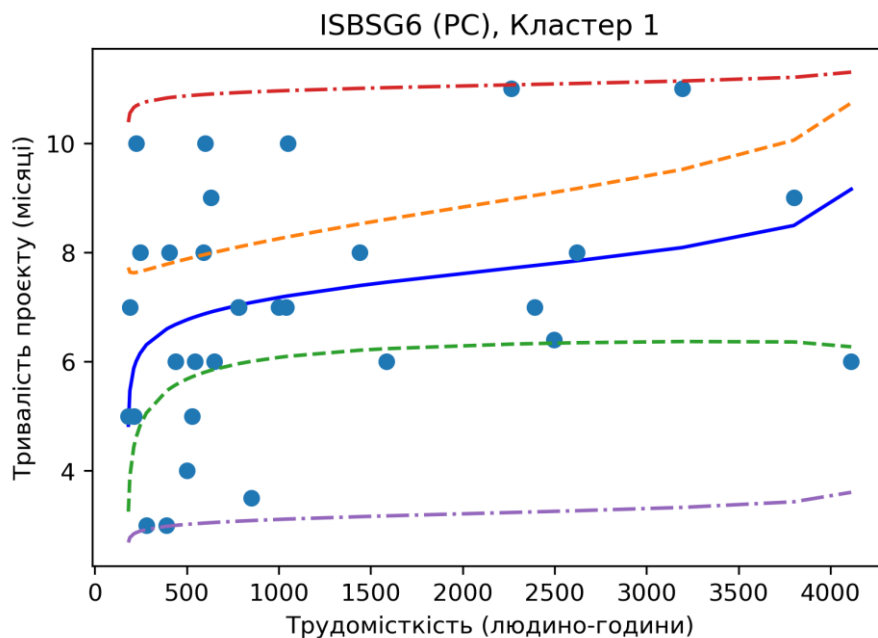


Рисунок 3 – Регресійна модель тривалості проєктів ПЗ, побудована з використанням одновимірного нормалізуючого перетворення Джонсона сімейства S_B

Для регресійної моделі тривалості проєктів ПЗ, побудованої з використанням нормалізуючого перетворення на основі десяткового логарифму, отримано значення $MMRE = 0,433$ та $Pred(0,25) = 0,333$. Для регресійної моделі тривалості проєктів ПЗ,

побудованої з використанням одновимірного нормалізуючого перетворення Джонсона сімейства S_B , отримано значення $MMRE = 0,285$ та $Pred(0,25) = 0,645$.

Висновки. Проведене дослідження показало, що є такі набори даних з програмних метрик, які можна розділити на окремі кластери та побудувати регресіні моделі для кожного з кластерів окремо. А при використанні одновимірного нормалізуючого перетворення Джонсона сімейства S_B отримати кращі показники якості у порівнянні з використанням нормалізуючого перетворення на основі десятичного логарифму.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- [1] Prykhodko, S. B., Prykhodko, N. V. (2023). *Building Nonlinear Regression Models for Estimating the Number of Clusters and Their Initial Centroids IEEE, 18th International Conference on Computer Sciences and Information Technologies (CSIT)*, 2023.
- [2] Rogel-Salazar, J. (2023, July 21). Getting started with K-means clustering in Python. <https://domino.ai/blog/getting-started-with-k-means-clustering-in-python>
- [3] Oligny, S., Bourque, P., Abran, A., Fournier, B. (2000). *Exploring the relation between effort and duration in software engineering projects, World Computer Congress, Aug. 2000.*

Prykhodko S.B.; Kudin O.O.

THE USE OF CLUSTER ANALYSIS METHODS IN THE CONSTRUCTION OF MODELS OF THE DURATION OF SOFTWARE DEVELOPMENT PROJECTS

Abstract. Nowadays, software developers use mathematical models of project execution time based on historical data. But they have unsatisfactory quality parameters. To improve the quality of the specified models, it is proposed to divide the general data set into clusters before building the model and build models for each of them.

Keywords: regression analysis; regression models of project time; cluster analysis; K-Means method; iterative clustering methods.