

ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ АНОМАЛІЙ В ДАНИХ

*д-р техн. наук, проф. С.Ю. Гавриленко, бакалавр В.Д. Зозуля,
Національний технічний університет "Харківський політехнічний
інститут", м. Харків*

Попередня обробка даних (preprocessing) направлена на підвищення якості самих даних та якості їх аналізу. Вона є нетривіальним завданням і може становити до 80% загального обсягу зусиль з інтелектуального аналізу даних.

Одним із важливих завдань очищення даних є виявлення аномалій. На сьогодні існує багато методів виявлення аномалій [1]. Але їх ефективність залежить від даних та параметрів і має слабкі систематичні переваги [2,3]. Окрім цього методи виявлення аномалій чутливі до параметрів налаштування моделей.

У доповіді досліджено такі методи виявлення аномалій: метод стандартного відхилення (Standard Deviation Method), метод локального рівня викидів (Local Outlier Factor), метод Ізольюючого лісу (Isolation Forest). Оцінка прийняття рішення щодо віднесення об'єкту до аномалій відбувалося за двома алгоритмами. на розмічених та нерозмічених даних. Отримано залежність кількості аномалій від порогу прийняття рішень для кожного із методів. Оцінку якості попередньої обробки даних виконано з використанням класифікаторів на основі методів KNN та беггінгу (Bagging).

За результатами дослідження більш якісним виявився алгоритм попередньої обробки даних Isolation Forest. Тестування показало збільшення точності класифікації за рахунок видалення аномалій для методу KNN – до 11,4%, для беггінгу – до 11,3%.

Досліджені методи реалізовані програмно з використанням хмарного сервісу GOOGLE COLAB на основі Jupyter Notebook

Проведені експерименти підтвердили працездатність методу Isolation Forest., що надає можливість рекомендувати його для практичного використання на етапі попередньої обробки даних з метою підвищення їх інтелектуального аналізу.

Список літератури: 1. Cui Z.G Research on preprocessing technology of building energy consumption monitoring data based on machine learning algorithm / Z.G. Cui, Y.Wu Cao, H.N. Liu, Z.F. Qiu, C. Chen Build. Sci. 2018, Vol. 34 (2), P. 94–99. 2. Borisyak Maxim (1+epsilon)-class Classification: an Anomaly Detection Method for Highly Imbalanced or Incomplete Data Sets / Maxim Borisyak, Artem Ryzhikov, Andrey Ustyuzhanin et al. // Journal of Machine Learning Research, 2020, Vol. 21(72), P. 1–22. 3. Гавриленко С.Ю. Розробка методу ідентифікації стану комп'ютерної системи на основі алгоритму "Isolation Forest" / С.Ю. Гавриленко, І.В. Шевєрдин // Радіоелектроніка, інформатика, управління, 2021, №.1(56), P. 105-116.