

ДОСЛІДЖЕННЯ МЕТОДІВ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ ДЛЯ ПІДВИЩЕННЯ ЯКОСТІ КЛАСИФІКАЦІЇ

Гавриленко С.Ю., Купін Д.О.

***Національний технічний університет
«Харківський політехнічний інститут», м. Харків***

Робота з даними неможлива без етапу їх попередньої обробки, яка включає набір процедур які направлені на підвищення якості самих даних та якості їх інтелектуального аналізу.

Виділяють наступні етапи попередньої обробки: очищення (Data cleaning), масштабування або нормалізація (Data transformation), перетворення даних (Scaling to a specific range), Скорочення або ущільнення даних (Data reduction), Секціонування даних (Data partitioning).

В даній роботі досліджено методи скорочення даних з використання хмарного середовище Google Colab Python.

Скорочення даних застосовується шляхом вибору об'єктів і атрибутів з метою спрощення обробки даних, зниження обчислювальних витрат, зменшення ймовірності перенавчання; збільшення семантичного розуміння моделі.

У якості вихідних даних використані дані про пацієнтів, які мають кардіологічні захворювання.

Розглянуто наступні методи скорочення даних: методи фільтри (Information Gain, Mean Absolute Difference, Variance Threshold), вбудовані методи (логістична регресія, дерева рішень на основі градієнтного спуску, Random Forest на основі мета-алгоритму беггінгу), методи обгортки (Wrapper methods).

Отримано, що найбільш ефективними є методи обгортки, які запускаються на різних підмножинах ознак вихідного тренувального сету. Процес вибору ознак базується на певному алгоритмі машинного навчання та дотримується підходу жадібного пошуку шляхом оцінки всіх можливих комбінацій ознак за заданим критерієм оцінки.

Література:

1. Kelleher J. , Namee B. and Archi A. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples and Case Studies ,The MIT Pres, 2015, 642 p.
2. Gavrylenko S. Y., Sheverdin I. V., “Development Of Method To Identify The Computer System State Based On The «Isolation Forest» Algorithm”, Radio Electronics, Computer Science, Control, 2021(1), pp. 105–116. doi:10.15588/1607-3274-2021-1-11
3. Harshita Singh. Understanding Data Preprocessing, 2020, URL: <https://towardsdatascience.com/data-preprocessing-e2b0bed4c7fb>
4. Gavrylenko S and Chelak V. Method of computer system state identification based on boosting ensemble with special preprocessing procedure, Advanced Information Systems, 2022. vol. 6, No 1, pp. 12–18. doi: DOI:10.20998/2522-9052.2022.1.02