ПОРІВНЯННЯ АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ K-СЕРЕДНІХ І C-СЕРЕДНІХ З ВИРІШЕННЯМ ПРОБЛЕМИ ВІДНОВЛЕННЯ ПРОПУЩЕНИХ ЗНАЧЕНЬ

Бородай Руслан Русланович, студент групи 335а*
Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ»

Кластеризація – це поділ множини вхідних векторів на групи (кластери) за ступенем «схожості» один на одного. Для того, щоб порівнювати об'єкти, необхідно мати критерій, на підставі якого буде відбуватися порівняння. Як правило, таким критерієм є відстань між об'єктами.

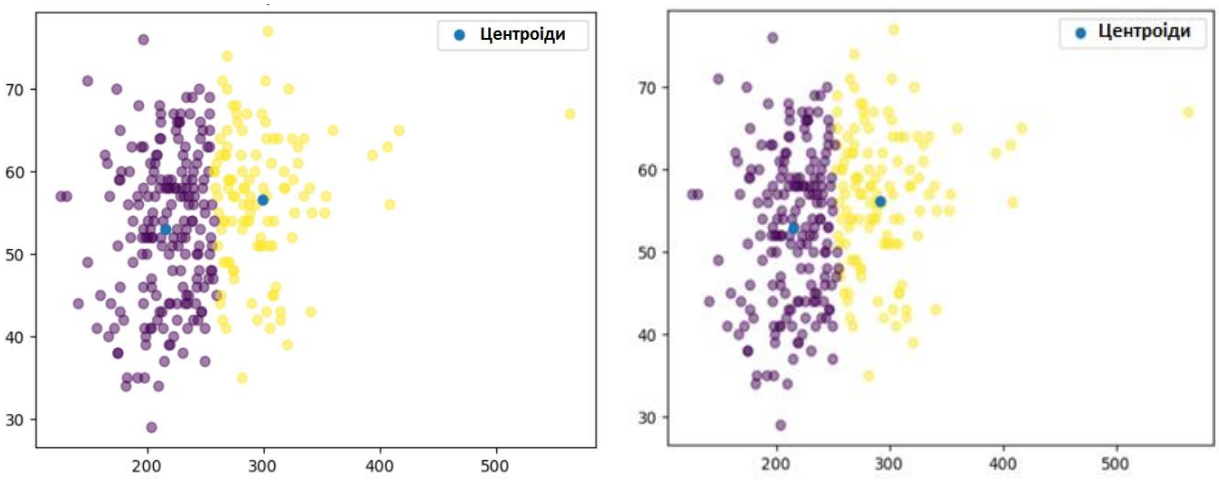
Алгоритм k-середніх – впорядкування множини об'єктів в порівняно однорідні групи [1-2]. Мета методу – розділити n спостережень на k кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції. Перевагами: простота та швидкість виконання, більш зручний для кластеризації великої кількості спостережень. Недоліки: результат класифікації залежить від випадкових початкових позицій кластерних центрів, алгоритм чутливий до викидів, які можуть викривлювати середнє, кількість кластерів повинна бути заздалегідь визначена дослідником.

Алгоритм нечіткої кластеризації (метод c-середніх) називають FCM-алгоритмом (Fuzzy Classifier Means, Fuzzy C-Means) [1-2]. Метою алгоритму кластеризації є автоматична класифікація множини об'єктів, які задаються векторами ознак у просторі ознак. Алгоритм визначає кластери і відповідно класифікує об'єкти. Кластери представляються нечіткими множинами, і, крім того, межі між кластерами також є нечіткими. FCM-алгоритм кластеризації припускає, що об'єкти належать усім кластерам з певною функцією приналежності. Ступінь приналежності визначається відстанню від об'єкта до відповідних кластерних центрів. Даний алгоритм ітераційно обчислює центри кластерів і нові ступені приналежності об'єктів. Основна перевага – визначення ймовірності того, що об'єкт належить до того чи іншого кластеру. Недоліки такі, як у k-середніх, але завдяки нечіткому розбиттю вони не є суттєвими.

Відновлення даних в даній програмі відбувається за допомогою інтерполяції. Одним з найкращих методів інтерполяції є імпутація простим середнім значенням, де пропущене значення заміщується середньоарифметичним значенням всього стовбця для цього використано бібліотеку «scikit-learn», клас «SimpleImputer». Вхідні дані отримуються у вигляді медичної викладки (файлу.xlsx або .xls). Для створення графічного інтерфейсу використовується бібліотека «Tkinter». Для створення графіків

використовується бібліотека «Matplotlib». Для роботи з даними використовується бібліотека «Pandas».

У роботі використано базу даних CardiologyCategorical.xls [3]. Кожен приклад представляє окремих пацієнтів та їхні різні медичні характеристики разом із класифікацією діабету. Кількість екземплярів: 303. Кількість атрибутів – 14.



(а) цілістні дані (б) з відновленими даними

Рисунок 1 – Графік c-means кластеризації

В результаті роботи можливо зробити висновок – помітна невелика розбіжність між алгоритмами k-means і c-means в центрі графіку (рис. 1) на даних, що були відновлені. Тобто результати кластеризації на вихідних даних та з використанням відновлених даних (1% від всіх значень стовпця) відрізняються не суттєво.

Список використаної літератури

1. Raschka, Sebastian, and Vahid Mirjalili. Python Machine Learning, 3rd Ed. Packt Publishing, 2019. 770 p. ISBN 1789955750.

2. Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, Edition 2. O'Reilly Media, Inc. 2019. 856 p. ISBN 9781492032595.

3. Supplemental Excel Data Sets [Електронний ресурс] – Режим доступу до ресурсу: http://mercury.webster.edu/aleshunas/Data%20Sets/Supplemental%20Excel%20Data%20Sets.htm.

**Виконано в рамках проєкту Національного фонду досліджень України 2020.02/0404 «Розробка інтелектуальних технологій оцінки епідемічної ситуації для підтримки прийняття управлінських рішень у сфері біобезпеки населення».*