

КЛАСТЕРИЗАЦІЯ ДАНИХ ЧАСОВИХ РЯДІВ: ОПТИМІЗАЦІЯ ПРОЦЕСУ КЛАСТИРИЗАЦІЇ ДЛЯ ТЕХНОЛОГІЧНИХ ЗМІННИХ

Для технологічних процесів виробництв та у дослідженнях даних соціальних, міських та інших сфер, виникає потреба знаходження та групування закономірностей у наборах експертних даних або даних моніторингу. В аналізі даних технологічних виробництв використовують стандартні методи на кшталт методи квадратичних помилок, моди, медіани, та K-mean, що в свою чергу не завжди можуть оцінювати дані процесів пов'язані між собою (корельовані). Для того щоб дослідити чи є залежності між змінними використовують ряд методів, один з яких буде розглянуто надалі.

При навчання з учителем, кількість класів у певному наборі даних відомо явно, оскільки кожен екземпляр даних записується як член одного з існуючих класів. У гіршому випадку ми можемо аналізувати атрибут класу і підрахувати кількість унікальних запитів, що існують. Для вирішення задачі кластеризації можна використовувати методи, один з яких подано нижче. Так званий метод “Індексів силуетів” оцінює подібність об'єкта до власного, створеного в процесі обробки інформації – кластера, цей процес здійснюють порівнюючи з іншими кластерами у процесі їх визначення. Інтерпретуючі знання методу - значення силуету є засобом порівняння, з діапазоном $[-1, 1]$. Кластеризований набір даних моделі, що видає переважно великі значення силуетів, швидше за все, є прийнятною та відповідною моделлю[1]. У нашій роботі використовувався програмний засіб MatLab з пакетом для роботи зі статистикою, до якого включено можливість роботи з індексами силуетів. При реалізації за допомогою цього пакету та визначивши попередньо кластеризацію методом середніх (K-mean) було отримано наступні результати. Коли кількість кластерів дорівнює 4 (для варіанту з часовими рядам витрат дифузійного соку та бурякової стружки), всі ділянки більше або менше мають аналогічну товщину і, отже, мають аналогічні розміри. Значення індексу силуету для кожної точки є мірою того, наскільки подібна ця точка до точок у своєму кластері, порівняно з точками в інших кластерах. За замовчуванням силует використовує квадратичну евклідову відстань між точками в X, що і показує Рис.1, на якому показано кількість класів (вісь Y) та розподіл існування відповідностей кластерам (вісь X).

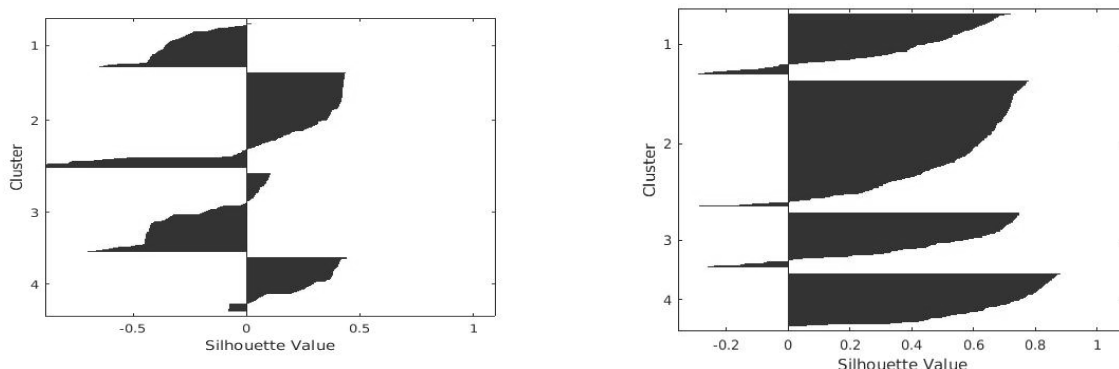


Рис.1 а - Силуети використовуючи квадратичну евклідову відстань між точками; б - тенденція кластерів силуетами за методом “Суми абсолютних різниць”

На Рис.1а показано, що за допомогою стандартної функції описати достатньо точно кластери неможливо бо існують великі розкиди початкових даних, які часто виникають у виробничих процесах через свою нестабільність та нестационарність, натомість можна

використати інший метод силуетів з тією самою кількістю кластерів. Для другого варіанту з 4ма кластерами отримуємо наступний результат Рис.1б. З вигляду й індексу силуету можна побачити, що більшість точок у другому кластері мають велике значення силуету більше 0.6, і що цей кластер має найбільше значення, що вказує на те, що кластер дещо відділений від сусідніх кластерів. В кожному кластері є незначна кількість від'ємно розміщених точок, що показує їхню погану роздільність, що не дозволяє точно їх призначити класу, але велика кількість точок має значення силуетів більше 0.7-0.8, що показує якісне їх віднесення до відповідних кластерів, що в свою чергу дає можливість покращити процеси що необхідно досліджувати.

Використання для покращення процесу ідентифікації методів таких як - метод Індексів силуетів дозволить ідентифікувати робочі режими роботи обладнання, тенденції, що виникають у економічних чи суспільних процесах та ряд інших залежностей. У роботі застосовані часові ряди, що взяті з реальних технологічних об'єктів та можуть у подальшому бути використаними для вирішення певних виробничих задач, що базуються на експертних та статистичних даних. Перевага методу Індексу силуетів у швидкому розрахунку та меншою дисперсією у порівнянні з великою кількістю інших статистичних методів.

Список використаних джерел

1. L. Kaufman. Finding Groups in Data: An Introduction to Cluster Analysis. / L. Kaufman, P. J. Rousseeuw. – MA: MIT Press, 2007 – pp. 215 – 266.

Науковий керівник: Ладанюк А. П., д.т.н., професор кафедри Автоматизації та інтелектуальних систем керування, Національний університет харчових технологій