УДК 004.94
Терещенко Е.В.[1] , Широкорад Д.В.[1], Рябенко А.Є.[1] , Царенко Є.С. [2]
[1] доц. НУ «Запорізька політехніка»
[2] студ. гр. КНТ-819сп НУ «Запорізька політехніка»

СТВОРЕННЯ ПРОГРАМНОГО МОДУЛЯ ДЛЯ КЛАСТЕРИЗАЦІЇ ДАНИХ МЕТОДОМ ІНКРЕМЕНТАЛЬНИХ СФЕР

Методи та алгоритми кластерного аналізу, як відомо, застосовуються у багатьох наукових дослідженнях з медицини, психології, археології та ін.

У науковій роботі [1] було запропоновано метод багатовимірної кластерізації (метод інкрементальних сфер), який має можливість визначати оптимальну кількість кластерів для заданої структури даних та може на відміну, наприклад, від поширеного методу k-середніх, бути застосований на складних структурах, таких як концентричні кола та ін.

З метою програмної реалізації методу був розроблений модуль IncSpheres з використанням мови програмування Python, та її бібліотек для обчислення та візуалізації даних NumPy, Matplotlib та Seaborn. Вибір мови програмування Python для розробки обумовлений тим, що вона є однією з лідируючих мов науки про дані. Наприклад, у бібліотеці NumPy представлені різноманітні та зручні інструменти для роботи з багатовимірними масивами, що дозволяють дуже швидко реалізовувати різноманітні обчислювальні процедури з мінімальною кількістю коду та у багатьох випадках без використання циклів. Бібліотеки Matplotlib та Seaborn дозволяють створювати та гнучко налаштовувати багато видів 2D та 3D графіків та діаграм.

Програмний компонент модулю створений за методологією об'єктно-орієнтованого програмування, тобто у виді класу. Основні методи: fit (обчислює кластеризацію наданих даних у вигляді багатовимірного масиву за методом інкрементальних сфер), predict (повертає масив позначок, що відносять об'єкти до різних кластерів, що потім використовується для графічного представлення даних, див. рис. 1), draw_diagram (будує діаграму залежності кількості кластерів від радіусу сфер).

При автоматичному створенні діаграми кількості кластерів акцент візулізації робиться на представленні оптимальної кількості кластерів (ця "сходинка" акцентується контрастним кольором), що можна побачити на рис.2.

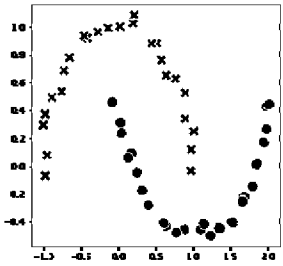

Рисунок 1 – Результат кластерного аналізу складної структури із застосуванням модуля IncSpheres.

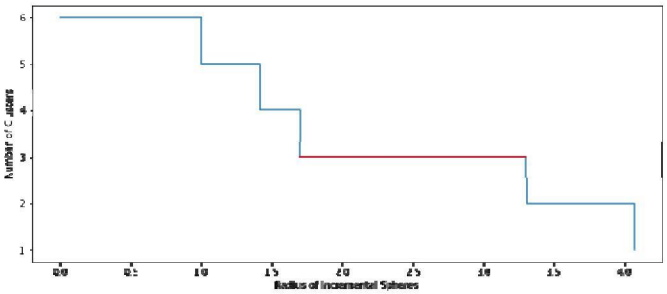

Рисунок 2 – Діаграма залежності кількості кластерів від радіусу інкрементальної сфери (оптимальна кількість кластерів - 3).

Модуль стандартизовано документований та готується до розміщення у глобальному репозиторії пакетів PyPI, після чого кожний користувач Python буде мати можливість його завантаження та застосування.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Пинчук В.П. Кластеризация данных методом инкрементальных сфер / В.П. Пинчук, А.Е. Рябенко // Системний аналіз. Інформатика. Управління (САІУ-2013) : матеріали IV Міжнародної науково-практичної конференції, [Запоріжжя], 13–16 березня 2013 року / Міністерство освіти і науки, молоді та спорту України, Академія наук вищої школи України, Запорізька обласна державна адміністрація, Класичний приватний університет. – Запоріжжя, 2013. – С. 206-207.