Технічні науки

УДК 336.72

Горносталь Олександр Миколайович
студент
НТУУ “Київський політехнічний інститут”
Горносталь Александр Николаевич
студент
НТУУ “Киевский политехнический институт”
Hornostal O.
student
NTUU “Kyiv Polytechnic Institute”

СИСТЕМА КЛАСТЕРНОГО АНАЛІЗУ ТА ВІЗУАЛІЗАЦІЇ ВЕЛИКИХ ОБСЯГІВ ДАНИХ
СИСТЕМА КЛАСТЕРНОГО АНАЛИЗА И ВИЗУАЛИЗАЦИИ БОЛЬШИХ ОБЪЕМОВ ДАННЫХ
SYSTEM OF CLUSTER ANALYSIS AND VISUALIZATION OF BIG DATA

Анотація: В роботі проведено огляд існуючих систем кластерного аналізу та візуалізації великих обсягів даних. Після дослідження та проектування було розроблено систему, призначену для: завантаження, нормалізації та аналізу вхідних даних; здійснення кластерного аналізу різними алгоритмами; візуалізації та збереження результатів проведеного кластерного аналізу.

Ключові слова: кластерний аналіз, аналіз даних, big data, кластер, міра відстані.

Аннотация: В работе проведен осмотр существующих решений систем кластерного анализа и визуализации больших объемов данных. После

исследования и проектирования была разработана система, предназначенная для: загрузки, нормализации и анализа входных данных; проведения кластерного анализа разными алгоритмами; визуализация и сохранение результатов проведенного кластерного анализа.

Ключевые слова: кластерный анализ, анализ данных, big data, кластер, мера расстояния.

Summary: In this work an overview of existed systems of cluster analysis and data visualization was made. After analysis and projecting was made system with such functionality: uploading, normalizing and analysis of input data; cluster analysis using various types of algorithms; visualization and saving results of cluster analysis.

Key words: cluster analysis, data analysis, big data, cluster, distance measure.

Вступ

Останнім часом інформація, що зростає в колосальних обсягах, породжує необхідність опрацювання великих обсягів даних. В цьому напрямку своє місце відведено інтелектуальному аналізу даних. Даний напрямок включає в кластерний аналіз та методи, основані на моделюванні, ймовірнісних узагальненнях, асоціюванні та пошуках закономірностей. Кластерний аналіз або автоматичне групування об’єктів є частковим випадком такого аналізу. В великій мірі розвитку цієї дисципліни сприяло проникнення в сферу аналізу даних ідей, що виникли в теорії штучного інтелекту.

Головним завданням кластерного аналізу є виділення необхідної кількості груп об’єктів, що об’єднані за певними критеріями між собою всередині групи і максимально відрізняються від екземплярів інших груп.

Аналіз математичного та алгоритмічного забезпечення систем кластерного аналізу

Кластерний аналіз групує об’єкти даних базуючись тільки на інформації, знайденій в даних, що описують об’єкти та їх взаємозв'язки. Мета в тому, щоб об’єкти в групі були схожими один до одного, але відрізнялися від об’єктів інших груп. Чим більша схожість об’єктів в групі, тим більше відрізняються групи, тим краща кластеризація[1].

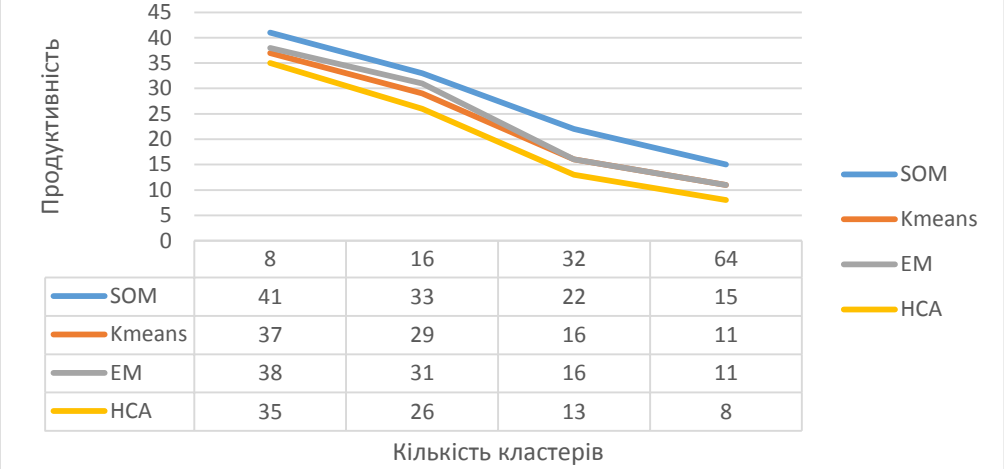
В роботі було обрано три прості, але важливі техніки для представлення багатьох концепцій в кластеному аналізі:

- к-середніх;

- агломеративна ієрархічна кластеризація;

- DBSCAN.

Проведено порівняння залежності продуктивності алгоритмів в залежності від кількості кластерів на наступних алгоритмах (рис. 1): к-середніх; алгоритм ієрархічної кластеризації; SOM (Self-Organization Map); EM (Expectation Maximization)[3].



	8	16	32	64
SOM	41	33	22	15
Kmeans	37	29	16	11
EM	38	31	16	11
HCA	35	26	13	8

Рис. 1 – Графік залежності продуктивності алгоритмів в залежності від кількості кластерів (розробка автора)

Зі зростанням кількості кластерів продуктивність SOM алгоритму падає. Продуктивність EM та К-середніх стає кращою ніж в ієрахічного алгоритму.

Кількість кластерів впливає також на якість кластеризації, тобто достовірність отриманих результатів.

Структура системи

Система кластерного аналізу складається з чотирьох основних блоків: блок аналізу, блок кластеризації, блок оцінки якості та інтерпретації кластерного рішення, засоби візуалізації [4]. Структура системи приведена на рисунку 2.

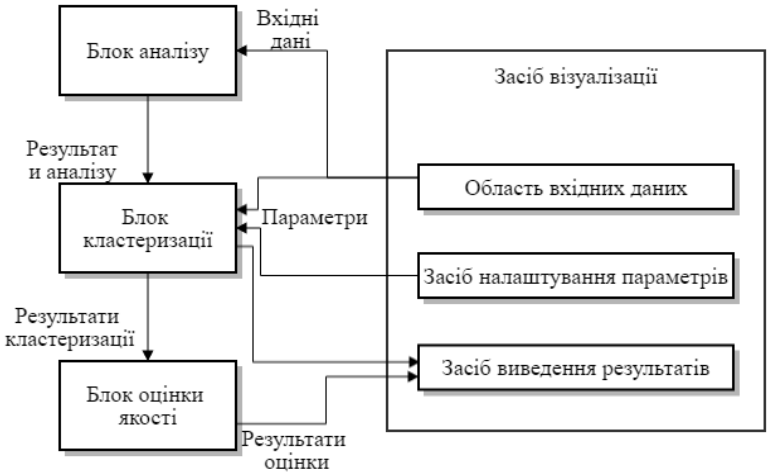


Рис. 2 – Структура системи (розробка автора)

Блок аналізу необхідний для виконання попереднього аналізу вхідних даних. Блок кластеризації використовується для безпосереднього розподілу об’єктів до кластерів. Планується реалізація різних типів алгоритмів кластерного аналізу даних. Засіб оцінки якості кластреризації призначений для оцінки степеня достовірності кластерних рішень, на основі яких будуть видані відповідні рекомендації. Засіб візуалізації представляє користувачу можливість взаємодії з системою.

Модель розробленого програмного забезпечення

Виходячи із побудованої концептуальної моделі була спроектована діаграма класів, що зображена на рисунку 3.

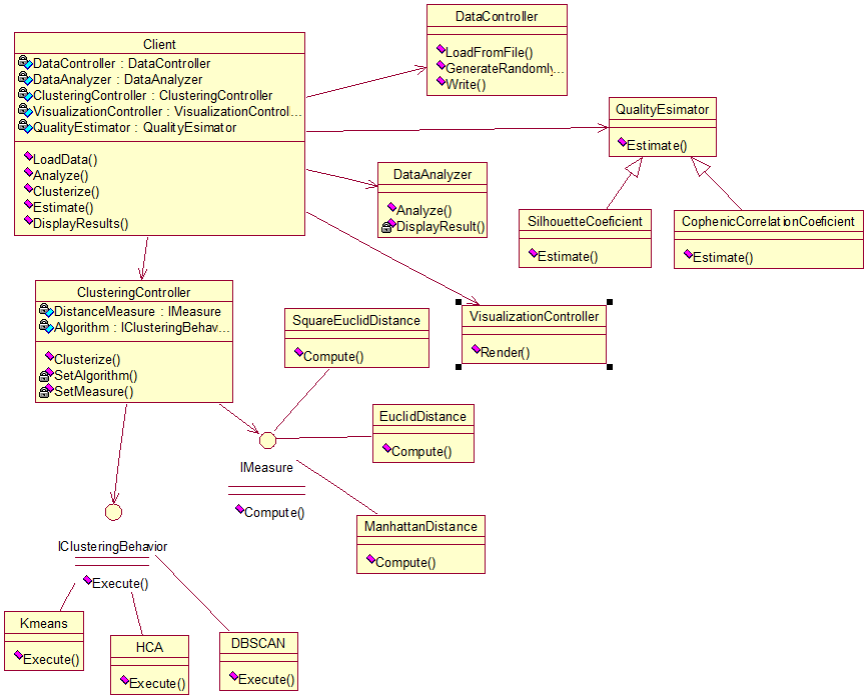


Рис. 3 – Діаграма класів системи (розробка автора)

Діаграма класів представляє логічне моделювання системи.

Результати роботи

Для демонстрації роботи програми застосуємо реалізований алгоритм кластеризації до набору даних, що можуть бути як згенеровані додатком автоматично, так і завантажені користувачем з файлу.

Набір даних містить кластери не сферичної форми з наявністю шумів в вигляді поодиноких точок. Такий набір кластерів найкраще виявляється алгоритмом DBSCAN при параметрах MinPts = 10, Eps = 4 (Рисунок 4).

Рис. 4 – Результати роботи алгоритму DBSCAN

Висновки

Було виконано дослідження предметної області, аналіз теоретичних засад та математичних методів кластерного аналізу, проектування, реалізацію та тестування програмного додатку системи кластерного аналізу та візуалізації великих обсягів даних.

З метою можливості подальшої підтримки програмного додатку код був написаний у відповідності до сучасних технологічних рішень, таких як паттерни проектування та з дотриманням SOLID принципів.

Розроблена система кластерного аналізу та візуалізації даних виконує поставлені перед нею задачі: аналіз даних, виконання алгоритмів кластеризаціі та візуалізація отриманих результатів.

Наступними напрямками розвитку системи можна визначити: реалізація більшої кількості алгоритмів кластеризації, розширення функціоналу для роботи не тільки з числовими, а й з категоріальними даними, покращення продуктивності при виконанні візуалізації.

Література:

1. A. K. Jain R. Algorithms for Clustering Data / R. C. Dubes A. K. Jain. – New Jersey: Prentice Hall, 1988. – 334 с.

2. Нейский И.М. Классификация и сравнение методов кластеризации [Електронний ресурс] / Нейский И.М. – Режим доступу до ресурсу: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf (дата звернення 18.05.2016). – Назва з екрану.

3. L. Kaufman. Finding Groups in Data: An Introdution to Cluster Analysis. Wiley Series in Probability and Statistics / L. Kaufman, P.J. Rousseeuw. – New York: John Wiley and Sons, 1990.

4. Програмная система кластерного анализа данных смешаного типа [Електронний ресурс] – Режим доступу до ресурсу: http://www.jurnal.nips.ru/sites/default/files/Paper-2013-1-11.pdf (дата звернення 30.05.2016). – Назва з екрану.