УДК 004.93'14

Р. Мельник, Р. Тушницький
Національний університет "Львівська політехніка",
кафедра програмного забезпечення

ДОСЛІДЖЕННЯ ПАРАМЕТРІВ ПРОСТОРУ ДАНИХ ПРИ ДВОКАСКАДНІЙ КЛАСТЕРИЗАЦІЇ

Для зменшення часових затрат при кластеризації даних великих розмірів запропоновано декомпозиційний підхід, що ґрунтується на розбитті простору за координатними осями гіперкубів. Відповідне керування алгоритмом дає змогу об'єднувати кластери – результати з підмножин – у кінцеві за незначними втратами точності. Як приклади практичних даних використані зображення із значними кількостями пікселів.

An approach to reduce algorithmic complexity for clustering of large-scale dataset is considered. The main idea is decomposition of item dataset and space by hypercube coordinates. To join clusters from subsets into the result clusters and to minimize the accuracy losses are the main tasks of the algorithm. Some visual patterns with large pixels numbers as test examples were investigated.

Вступ

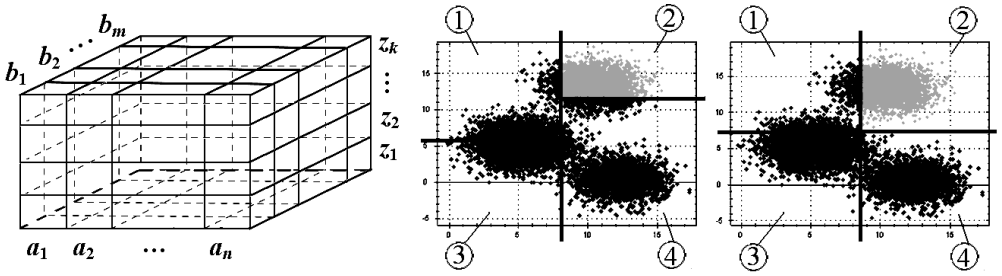
Методи кластерного аналізу широко використовуються для декомпозиції, дослідження, індексування та пошуку, розпізнавання зображень [1–4]. Зокрема, робота [1] містить класифікацію методів кластеризації та спосіб формування контурів виділених кластерів. Роботи [2–3] присвячені кластеризації графових моделей, якими відображають частини зображень. Ієрархічний алгоритм кластеризації в [4] має один етап процедури згортання і пропонує новий критерій об'єднання для зменшення обчислювальних затрат із складності $O(N^3)$ до $O(N^2)$. Для великих обсягів даних з багатьма характеристиками ця складність виявляється також перешкодою для групування. Тому в роботі пропонуються декілька підходів для зменшення складності на основі декомпозиції простору пошуку кластерів. Перші спроби продемонстровано в роботах [5, 6].

Декомпозиція простору

Практичні класи задач кластеризації даних є часоплинними завдяки об'ємам даних і складності алгоритму. Обчислювальна складність класичного ієрархічного алгоритму сягає значення $O(N^3)$, а певними кроками обмеження перебору можна досягти величини $O(N^2)$, що підтвердили експерименти роботи [4] з дослідження залежності часу побудови дерева згортання від величини початкової вибірки. У роботі [6] реалізовано наближений підхід кластеризації ключів з розбиттям множини базових ключів. Для подальшого зменшення складності алгоритму ієрархічної кластеризації пропонується декомпозиційний підхід, який ґрунтується на розбитті початкової вибірки даних великої розмірності на ряд підмножин. При цьому розглядаються дві можливості: розбиття вибірки на частини з виконанням обмежень на кількість значень у підмножині та розбиття простору на частини без обмежень на потужність підмножин значень даних.

Підхід з декомпозицією всього простору n-вимірних даних полягає в поділі всіх координат на частини і побудови гіперкубів. Фактично рознесення значень точок n-вимірного простору у відповідний гіперкуб вимагає попереднього опрацювання даних, а саме сортування значень вибірки за всіма координатами.

Нехай простір складається із s n-вимірних даних: $C = C(a, b, \ldots, z)$. Вимірність a поділимо на n частин, b – на m частин, ..., вимірність z – на k частин. Поділ простору за вимірностями на частини за параметрами $n, m, \ldots, k$ позначимо вектором розбиття $l = (n, m, \ldots, k)$. Схематично поділ тривимірного простору зображено на рис. 1, *а*.



а *б* *в*

Рис. 1. Схема поділу тривимірного простору на куби (а) та поділ двовимірного простору: б – гіперкуби із однаковою кількістю точок, в – за значеннями по кожній координаті

Розглядаємо два способи поділу простору: гіперкуби різних розмірів, але з однаковою кількістю значень дних; другий – на гіперкуби з неконтрольованими кількостями значенннями даних, але контрольованими розмірами сторін, що задаються користувачем. На рис. 1, *б–в* зображено простір, що складається із 20000 2-вимірних точок, згенерованих за нормальним законом розподілу. На рис. 1, *б* вектор розбиття l = (2, 2) дає гіперкуби (прямокутники у 2-вимірному просторі) із однаковою кількістю точок, але різних за розмірами. На рис. 1, *в* цей самий вектор розбиття та поділ координатних проміжків на однакові частини дає однакові гіперкуби. На рис. 1, *б* – кожна підмножина (кожен квадрант) – містить по 5000 точок. На рис. 1, *в* зображено: в 1-му квадранті 2030 точок, в 2-му – 11350, в 3-му – 4540, в 4-му – 2080 точок.

Складність алгоритму збільшується на сортування вибірки за всіма координатами, тобто стає рівною: $O(p \cdot n_i^3) + p \cdot O(N^2)$.

Алгоритм двокаскадної кластеризації

За базовий приймаємо класичний ієрархічний алгоритм згортанння – побудови дерева кластерів різних рівнів. В даній роботі пропонується багаторівнева декомпозиція множин ключів і кластерів, яку назвемо *каскадною кластеризацією*.

Розбиваємо вхідну множину елементів $Q(Q_1, Q_2, Q_3, \dots Q_N)$ на p підмножин $O_1(Q_1, Q_2, Q_3, \dots Q_z)$, $O_2(Q_{z+1}, Q_{z+2}, Q_{z+3}, \dots Q_t)$, … , $O_p(Q_{t+1}, Q_{t+2}, Q_{t+3}, \dots Q_N)$. До кожної з підмножин (назвемо їх множинами нульового каскаду, застосуємо алгоритм кластеризації, утворивши множини відповідних кластерів $K_1(k_1, k_2, k_3, \dots)$, $K_2(k_s, k_{s+1}, k_{s+2}, \dots)$, … , $K_p(k_r, k_{r+1}, k_{r+2}, \dots)$, де $k_1, k_2, \dots, k_i, \dots$ – кластери, елементи яких відносяться до відповідних підмножин $O_1, O_2, \dots, O_p$. Утворимо множину кластерів 1-го каскаду кластеризації K об'єднанням:

$$K = K_1(k_1, k_2, k_3, \dots) \cup K_2(k_s, k_{s+1}, k_{s+2}, \dots) \cup \dots \cup K_p(k_r, k_{r+1}, k_{r+2}, \dots). \quad (1)$$

Застосуємо до цієї множини алгоритм кластеризації, розглядаючи кожен з кластерів $k_1, k_2, \dots, k_i, \dots$ як базовий, тобто листок дерева згортання. В результаті утворюється множина кластерів 1-го каскаду. Тобто отримуємо двокаскадну декомпозицію простору. Схему поділу та згортання зображено на рис. 2.

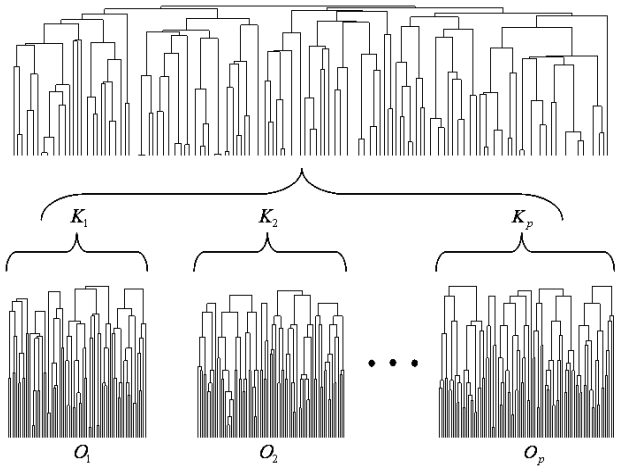


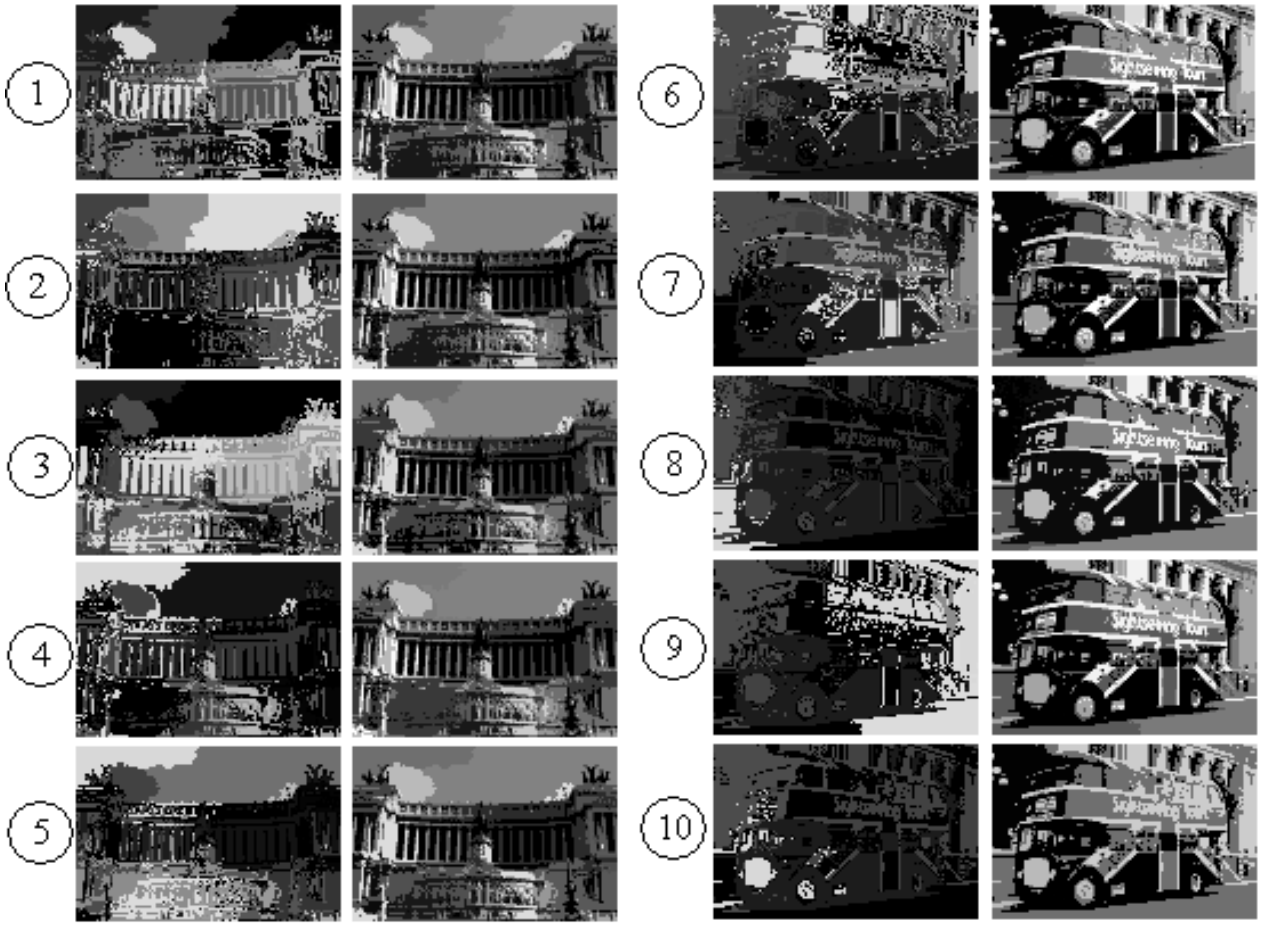
Рис. 2. Каскадне дерево формування кластерів

Керування вектором розбиття та кількістю кластерів 1-го каскаду

Проілюструємо результати роботи алгоритмів на зображеннях розмірами 116 × 77 пікселів. Вхідна вибірка – це координати *x*, *y* та *color* – середнє арифметичне кольорів пікселів. Загалом дані містять 8932 тривимірних точок.

У першому експерименті з підгрупи (гіперкуба) прийнято вихід кластерів, що дорівнює 10 % від їх кількості у цій підгрупі (кількість кластерів 1-го каскаду) та вибрано такі вектори розбиття *l*: (2, 2, 1), (2, 2, 2), (4, 4, 1), (4, 4, 2) та (4, 4, 4). У другому експерименті вибрано вектор розбиття *l* = (2, 2, 2) та кількість кластерів 1-го каскаду 2%, 5%, 10%, 15% та 20% від їх кількості у цій підгрупі.

На рис. 3 представлено результати декомпозиції простору та каскадного згортання першого і другого експериментів. Зокрема, на рис. 3, *а–б* наведено результуючі кластери, розфарбовані випадковим кольором та їх зваженою яскравістю для різних векторів розбиття. На рис. 3, *в–г* представлено результуючі кластери, розфарбовані випадковим кольором та їх зваженою яскравістю для різної кількості кластерів 1-го каскаду.



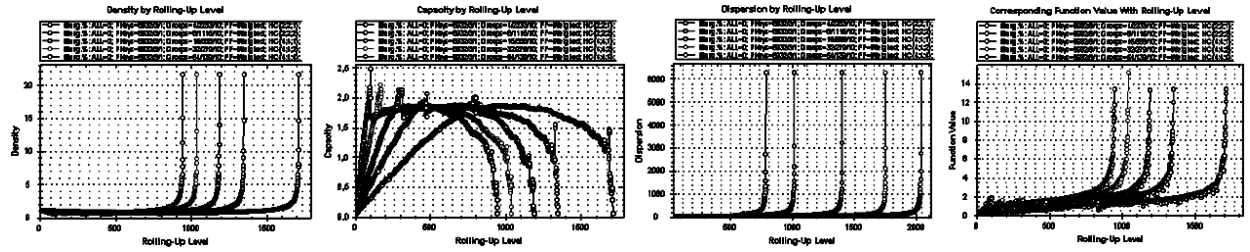
Кількість кластерів 1-го каскаду = 10% *Вектор розбиття l = (2, 2, 2)*

а *б* *в* *г*

Рис. 3. Результати декомпозиції простору та каскадного згортання: 1– l = (2, 2, 1); 2 – l = (2, 2, 2); 3 – l = (4, 4, 1); 4 – l = (4, 4, 2); 5 – l = (4, 4, 4); 6 – 2%; 7 – 5%; 8 – 10%; 9 – 15%; 10 – 20%; а, в – кластери, розфарбовані випадковим кольором; б, г – кластери, розфарбовані їх зваженою яскравістю

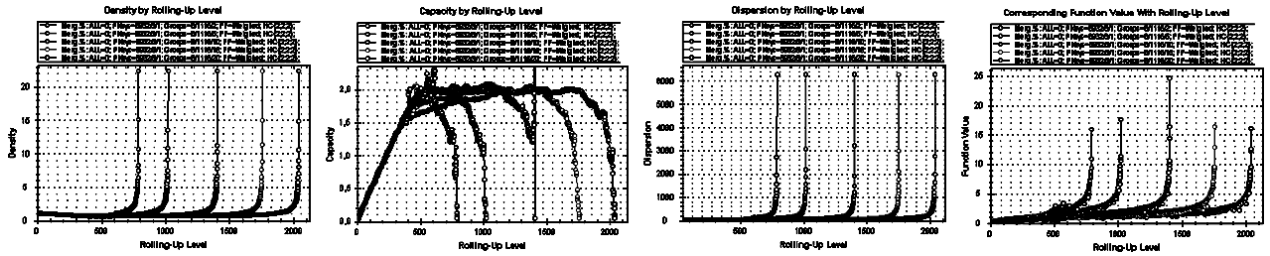
На рис. 4, 5 представлено графіки питомих функцій та функції сусідства каскадного згортання для даних експериментів. Графіки показують зміну питомої густини, питомого об’єму, питомої дисперсії та функції сусідства залежно від рівня дерева згортання та параметрів керування: кількість кластерів 1-го каскаду та вектора розбиття.

У табл. 1, 2 наведено числові характеристики кластеризації для першого та другого експериментів.



а *б* *в* *г*

Рис. 4. Графіки питомих функцій та функції сусідтсва каскадного згортання для кількості кластерів 1-го каскаду у 10%: а − питома густина; б − питомий об'єм; в − питома дисперсія; г − функція сусідства



а *б* *в* *г*

Рис. 5. Графіки питомих функцій та функції сусідтсва каскадного згортання для вектора розбиття l = (2, 2, 2): а − питома густина; б − питомий об'єм; в − питома дисперсія; г − функція сусідства

Таблиця 1

Характеристики кластеризації для різного вектора розбиття

Вектор розбиття	Час згортання, хв	Кількість кластерів в групі	Рівень (всього рівнів)	Кількість кластерів на рівні	Питома густина	Питомий об'єм	Питома дисперсія
(2,2,1)	2,08	2233	1693 (1716)	23	1,9778	1,4592	16,8139
(2,2,2)	0,44	1116	1330 (1352)	22	1,9726	1,4679	16,7161
(4,4,1)	0,21	558	1170 (1190)	20	1,9635	1,4838	17,1928
(4,4,2)	0,12	279	1020 (1041)	21	1,9617	1,4894	17,2382
(4,4,4)	0,09	139	927 (948)	21	1,9555	1,5036	17,5113

Таблиця 2

Характеристики кластеризації для різної кількості кластерів 1-го каскаду

К-ть кластерів 1-го каскаду, %	Час згортання, хв	К-ть кластерів 1-го каскаду	Рівень (всього рівнів)	Кількість кластерів на рівні	Питома густина	Питомий об'єм	Питома дисперсія
2	0,42	22	776 (792)	16	1,8675	1,6215	20,2668
5	0,43	55	1010 (1022)	12	1,8676	1,6169	20,3030
10	0,52	111	1400 (1407)	7	1,8728	1,6131	20,7958
15	1,15	167	1749 (1758)	9	1,8677	1,6073	19,6580
20	1,52	223	2038 (2043)	5	1,8707	1,6049	19,7236

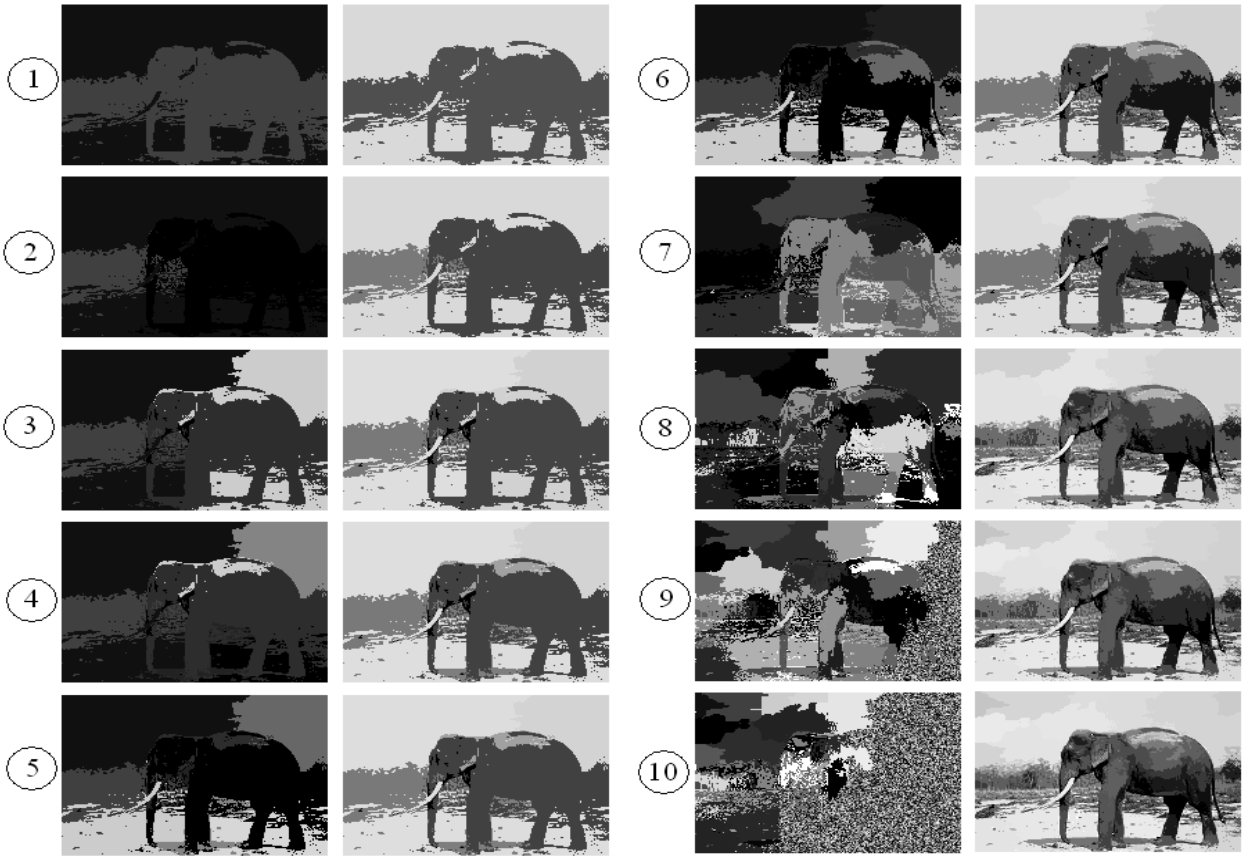
З даних табл. 1, 2 можна дійти таких висновків: розбиття простору на більшу кількість підпросторів або зменшення кількості кластерів 1-го каскаду дає змогу кластеризувати дані швидше. При цьому:

− питома густина зменшується, а питома дисперсія і питомий об'єм збільшується;

− кількість рівнів дерева згортання зменшується;

− необхідна кількість кластерів для опису загальної картини образу із збільшенням кількості підпросторів або кількості кластерів 1-го каскаду зменшується.

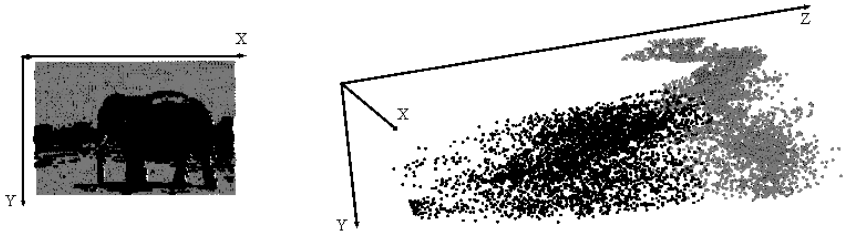
На рис. 6 представлено кластери, що знаходяться на різних рівнях ієрархічного дерева згортання, розфарбовані випадковим кольором та їх зваженою яскравістю.



а *б* *в* *г*

Рис. 6. Кластери на різних рівнях дерева згортання: 1 – рівень 1449, кластерів 2; 2 – рівень 1448, кластерів 3; 3 – рівень 1446, кластерів 5; 4 – рівень 1443, кластерів 8; 5 – рівень 1441, кластерів 10; 6 – рівень 1439, кластерів 12; 7 – рівень 1435, кластерів 16; 8 – рівень 1400, кластерів 51; 9 – рівень 1380, кластерів 71; 10 – рівень 1300, кластерів 151; а,в − кластери, розфарбовані випадковим кольором; б, г − кластери, розфарбовані їх зваженою яскравістю

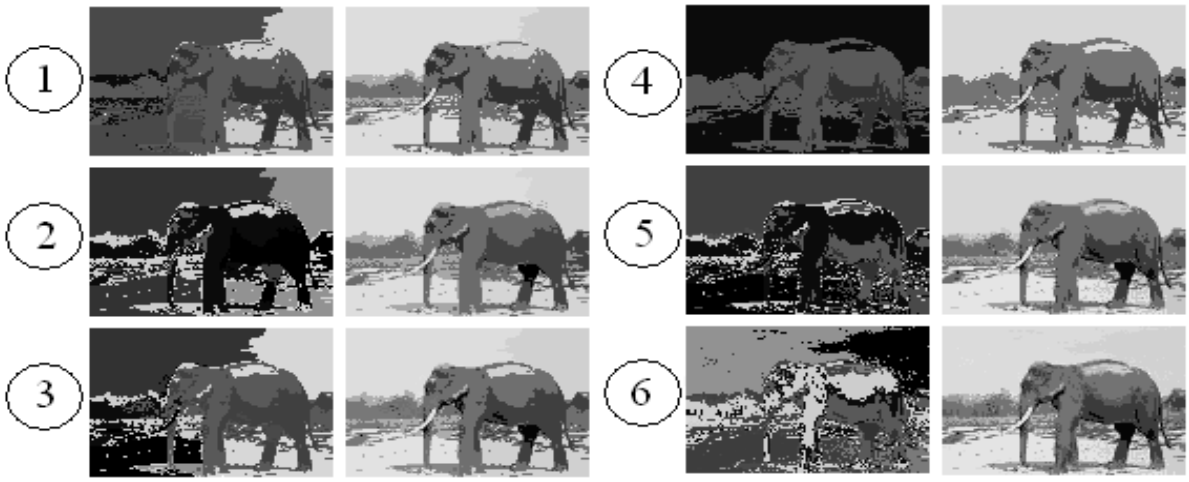
Для представлення каскадної декомпозиції простору наведемо порівняльні результати досліджень роботи алгоритмів для вхідних вибірок, що складаються із тривимірних точок (координати *x, y* та *color* – середнє арифметичне кольорів пікселів) та п'ятивимірних точок (координати *x, y* та компоненти *R, G, B* кольору пікселів), отриманих із тестових зображень. Разом дані містять 8932 елементів. На рис. 7 представлено 3 результуючі кластери декомпозиції тривимірних даних із зображення: *а* − осі координат відповідають координатам *x, y*; *б* − вісь *z* відповідає значенню *color*.



а *б*

Рис. 7. Декомпозиція простору на 3 кластери

На рис. 8 представлено результати декомпозиції 3- та 5-вимірних точок із різним вектором розбиття та кількістю кластерів 1-го каскаду 20%. Числові характеристики декомпозиції зведено у табл. 3.



тривимірні точки *п'ятивимірні точки*

а *б* *в* *г*

Рис. 8. Результати декомпозиції простору та каскадного згортання: 1– l = (2, 2, 2); 2 – l = (3, 3, 3); 3 – l = (4, 4, 4); 4 – l = (2, 2, 2, 2, 2); 5 – l = (2, 2, 4, 4, 4); 6 – l = (1,1, 9, 9, 9); а, в – кластери, розфарбовані випадковим кольором; б, г − кластери, розфарбовані їх зваженою яскравістю

Таблиця 3

Характеристики кластеризації три- та п'ятивимірних точок із різним вектором розбиття

Вектор розбиття	Час згортання, хв	Кількість кластерів в групі	Рівень (всього рівнів)	Кількість кластерів на рівні	Питома густина	Питомий об'єм	Питома дисперсія
(2,2,2)	1,46	1116	2034 (2040)	6	2,0839	1,3007	14,1585
(3,3,3)	1,15	330	1801 (1810)	9	2,0722	1,3116	14,3695
(4,4,4)	1,09	139	1724 (1734)	10	2,0615	1,3274	14,4147
(2,2,2,2,2)	1,24	279	1660 (1663)	3	1,6185	2,8027	34,5558
(2,2,4,4,4)	1,08	34	1509 (1518)	9	1,5621	3,1530	39,7618
(1,1,9,9,9)	0,55	12	1386 (1398)	12	1,3870	5,3030	116,7975

Висновки

Розроблено підхід двокаскадної кластеризації, що ґрунтується на розбитті простору та даних відповідно до нього на частини. Основну увагу приділено дослідженню параметрів простору даних: вектора розбиття та кількості кластерів 1-го каскаду. За цим методом можна опрацювати дані, які точним алгоритмом зробити неможливо. Втрати точності відбуваються на межах гіперкубів. Подальші дослідження необхідні для мінімізації зазначених втрат.

1. Andy M Yip, Chris Ding, Tony F.Chan. Dynamic Cluster Formation Using Level Set Methods // IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.28, n. 6, pp.877–889, June, 2006. 2. Leo Grady, Eric L.Schwartz. Isoperimetric Graph partitioning for Image segmentation // IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.28, n. 3, pp.469-475, March, 2006. 3. M Pavan, M Pelillo. Dominant sets and Pairwise Clustering // IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.29, n. 1, pp.167-172, January, 2007. 4. C. Ding, X. He. Cluster Aggregate Inequality and Multilevel Hierarchical Clustering // Proc. 9th European Conf. Principles of Data Mining and Knowledge Discovery, pp. 71−83, 2005. 5. R. Melnyk, R. Tushnytskyy. Algorithm Accuracy and Complexity Optimization by Inequality Merging for Data Clustering // Proc. of the Xth Intern. Conf. The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM'2009), pp. 453-455, 2009. 6. Мельник Р., Тушницький Р. Каскадна декомпозиція множин великої розмірності при кластеризації ключів образів // Комп'ютерні науки та інформаційні технології. – 2008. – № 604. – С.249–254.