Магістерська робота
Алгоритми кластеризації даних великих об'ємів

виконав Волощук О.Р.
керівник ас. Годич О.В.

16 червня 2011 р.

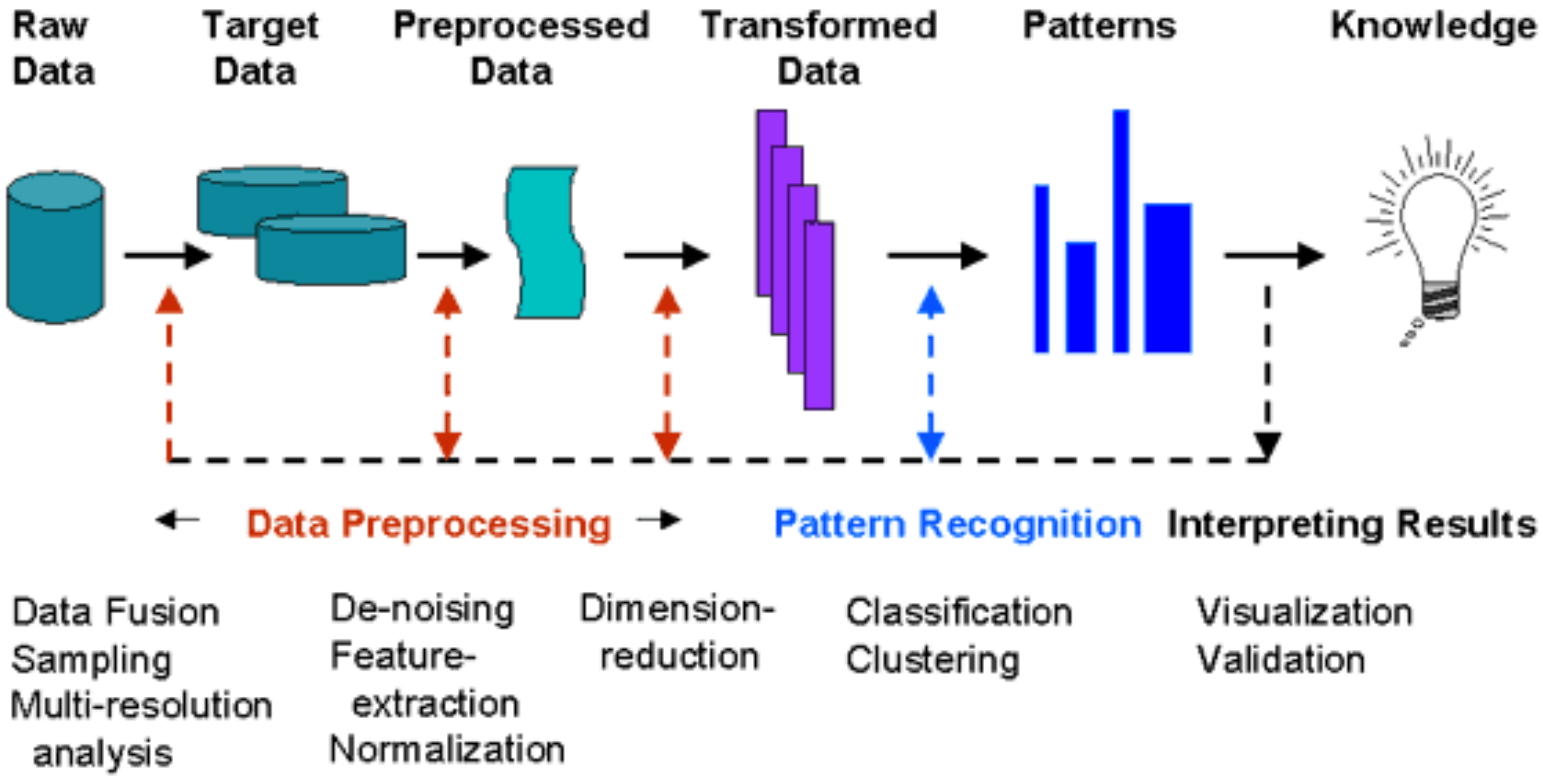
Задача кластеризації

Нехай D — множина точок n-вимірного простору.

Означення

Кластеризацією $C = \{C \mid C \subseteq D\}$ *називається таке розбиття* D *на підмножини, для якого виконується* $\cup_{C_i \in C} = D$ *і* $\forall C_i, C_j \in C : C_i \cap C_{j \neq i} = \emptyset$. *Множини* C_i *називаються кластерами.*

Процес видобування знань



Застосування

- розпізнавання зображень, мови
- соціологія
- медицина
- маркетингові дослідження

Алгоритми

- K-means
- DBSCAN
- UPGMA
- Neighbor-joining

Тестові дані

Тестування швидкодії алгоритмів проводилось на наборі даних розміром до 100000 об'єктів розмірності 8, 32 та 64.
Кожен об'єкт вибірки — вектор, всі компоненти якого лежать у проміжку $(-1; 1)$ та є випадковими величинами.

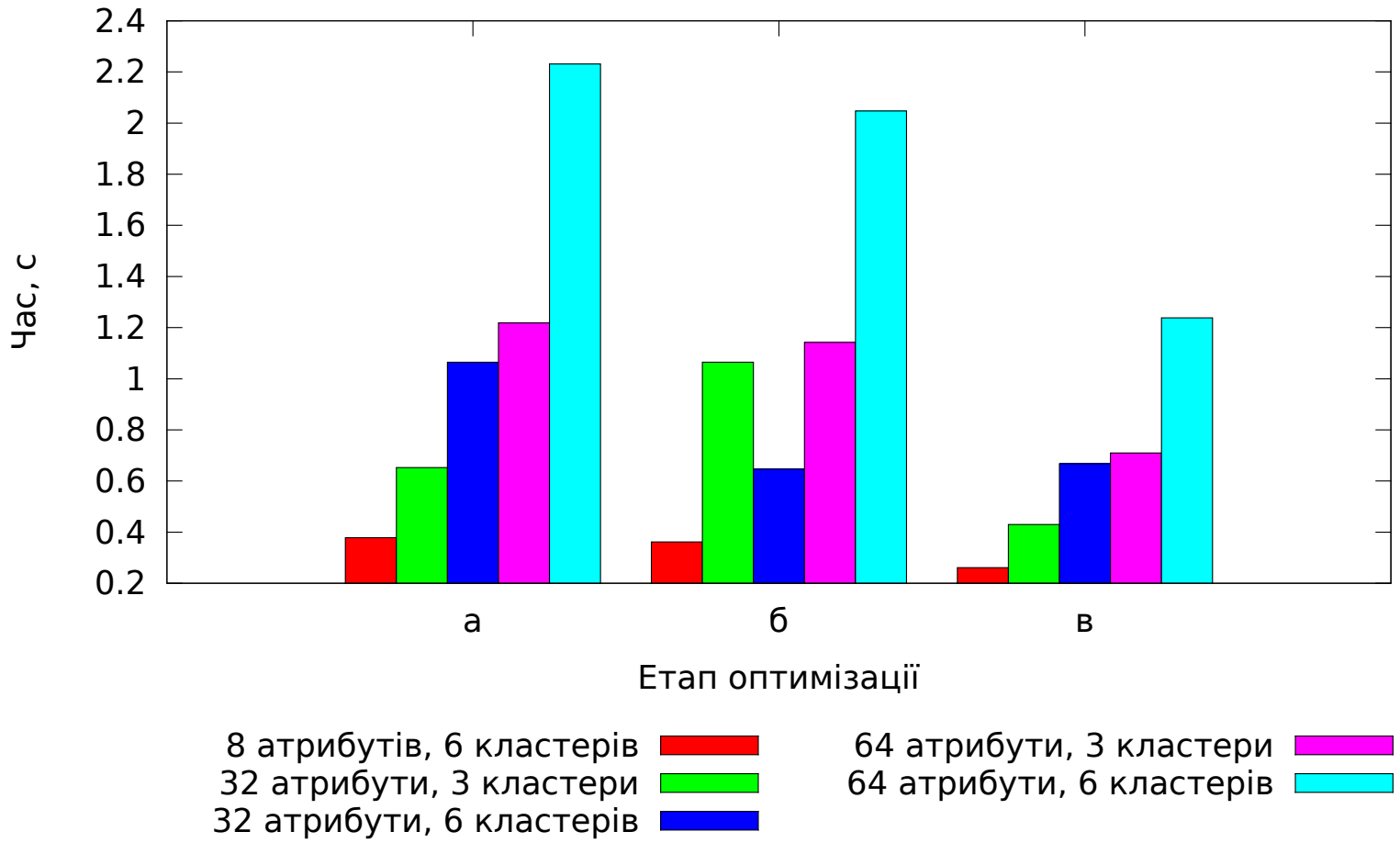
Критерії оцінки ефективності

Ефективнсть реалізації кожного алгоритму оцінювалась в першу чергу за часом роботи.
Для усіх алгоритмів час виконання одної ітерації не змінюється на протязі всього часу роботи, тому оцінювати можна зміни часу виконання ітерації.

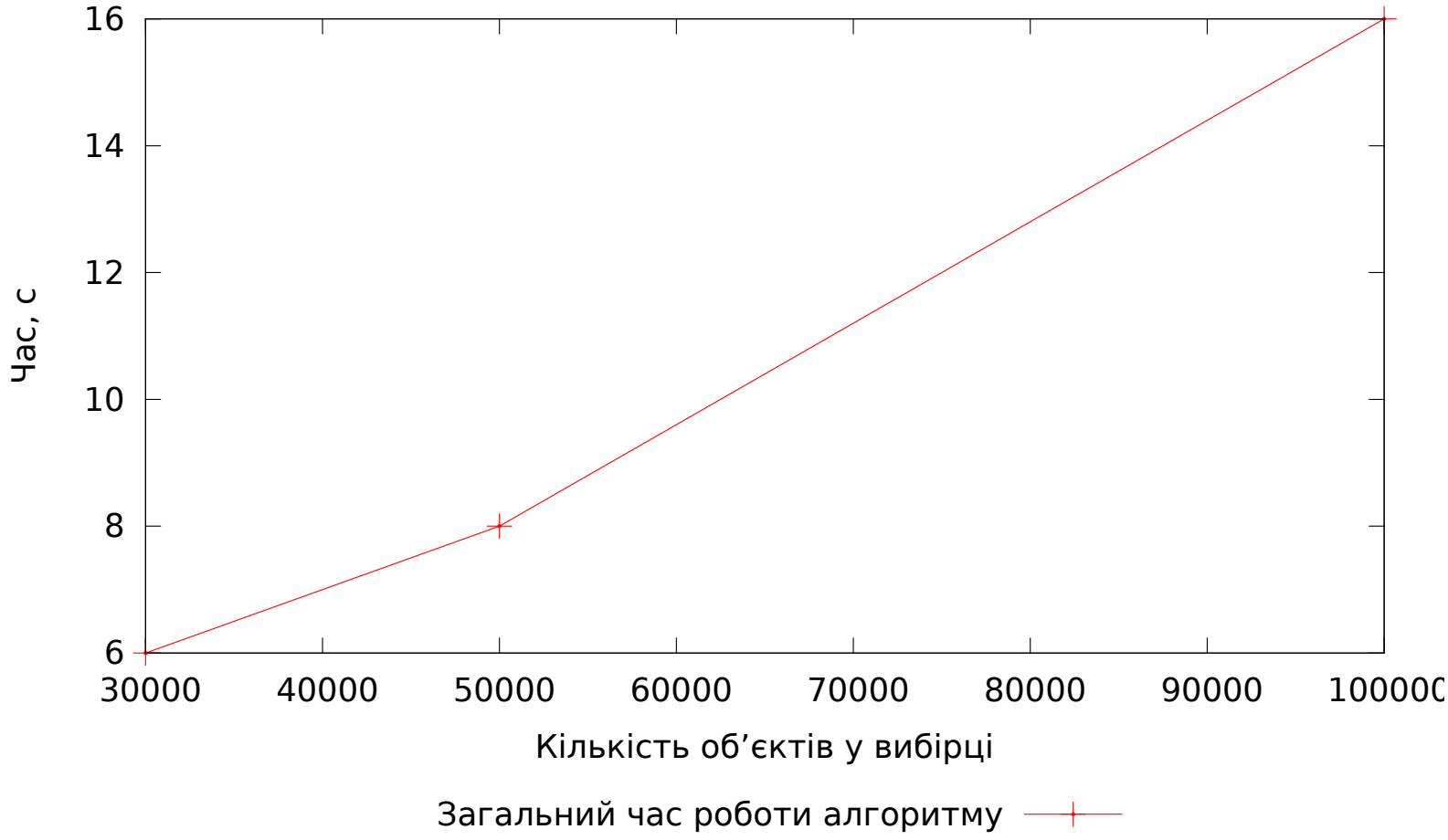
Програмне забезпечення

Для перевірки швидкодії алгоритмів створено програмну реалізацію кожного із них.
З міркувань швидкодії реалізацію створено за допомогою мови програмування C++, що дозволяє ефективно керувати пам'яттю та використовувати переваги багатопроцесорних архітектур.

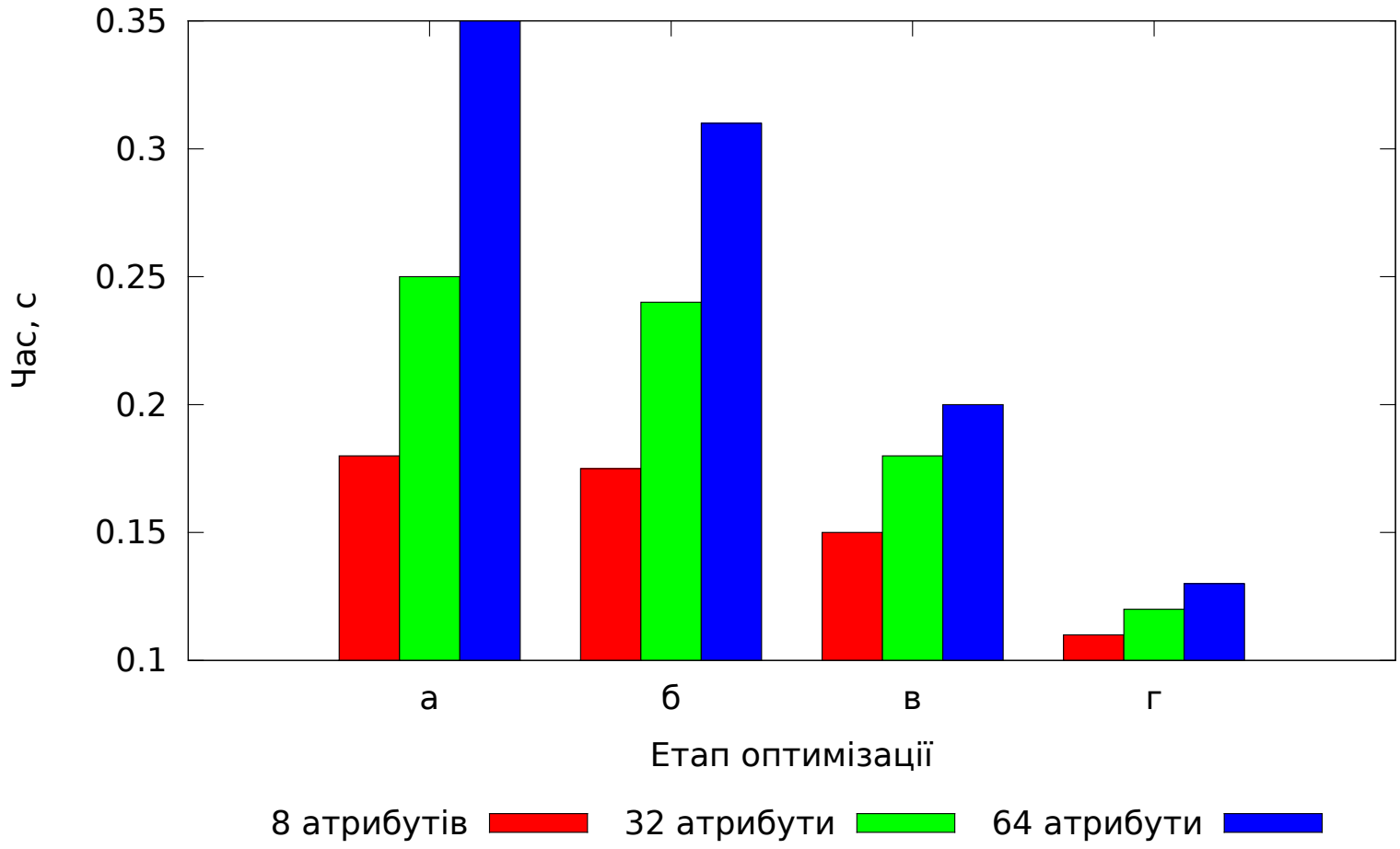
K-means — оптимізації



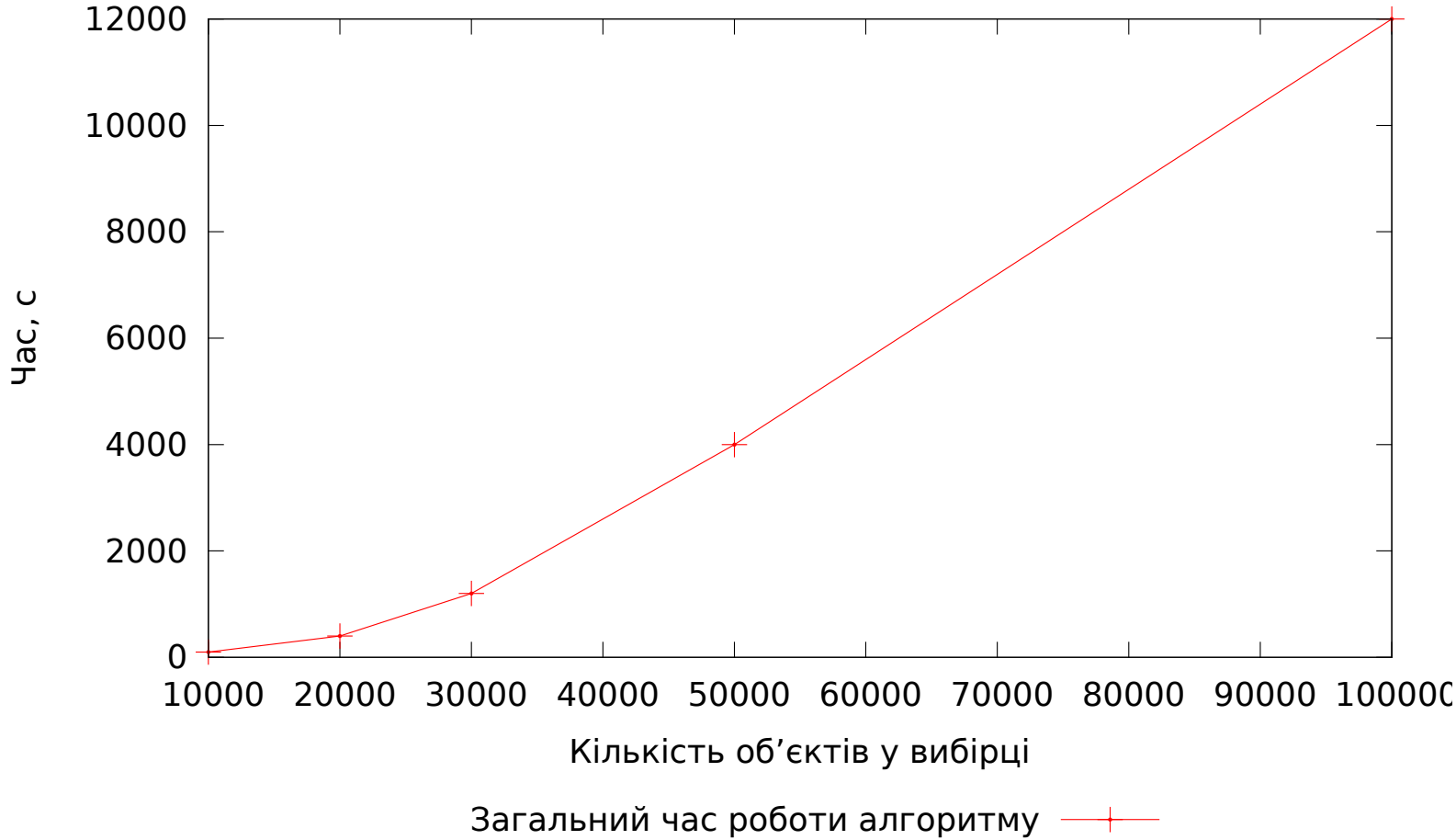
K-means — загальний час роботи



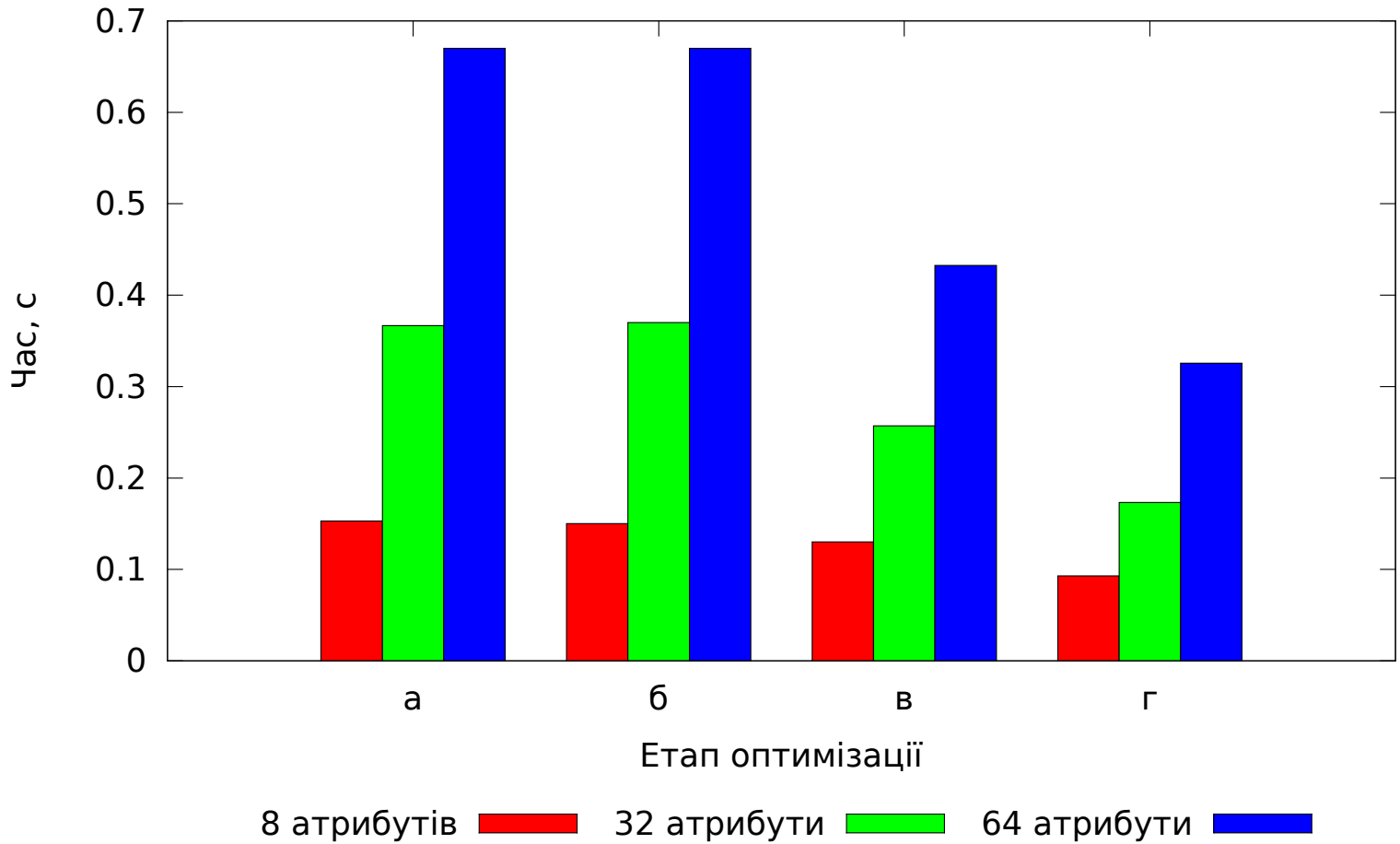
DBSCAN — оптимізації



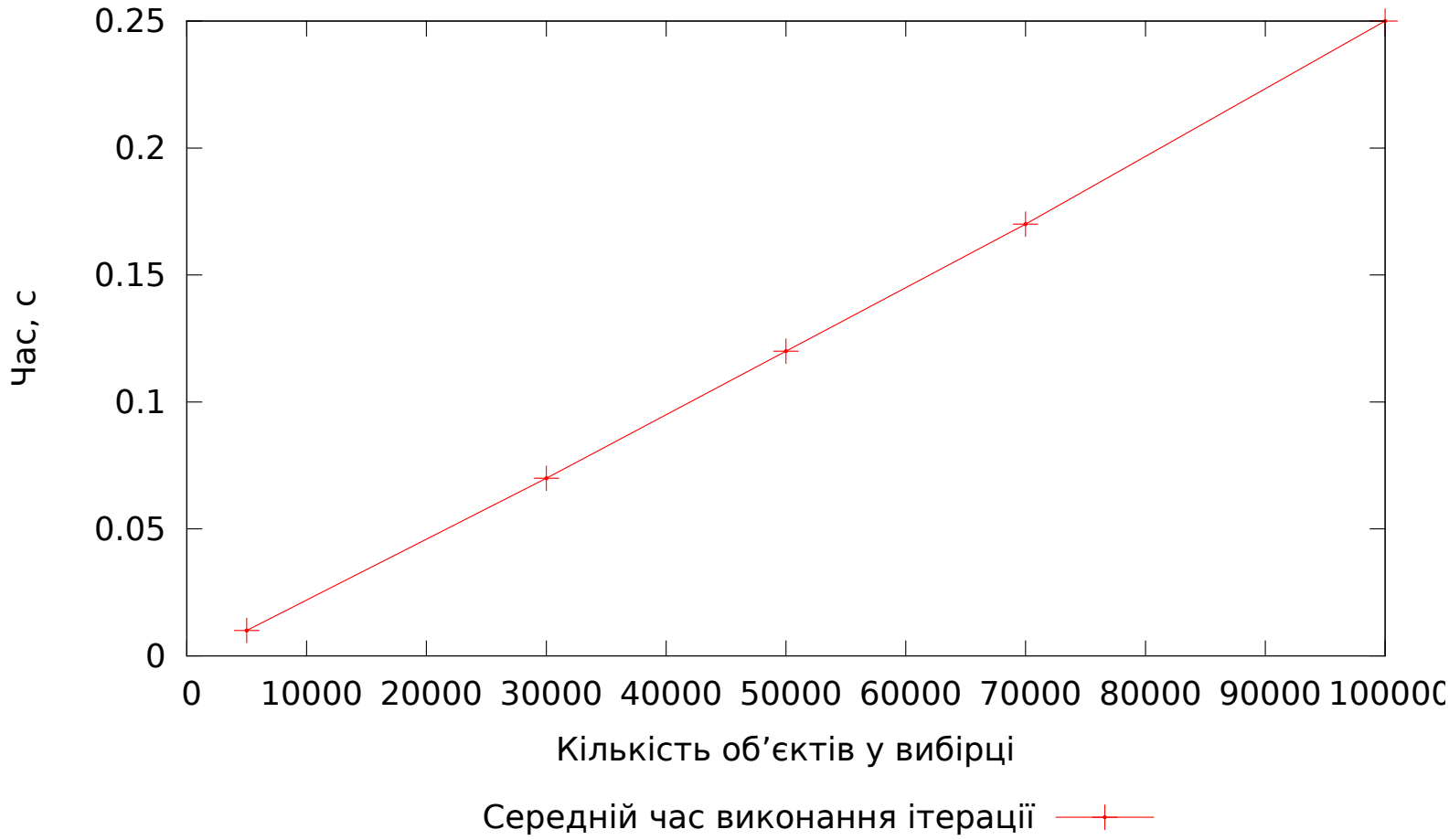
DBSCAN — загальний час роботи



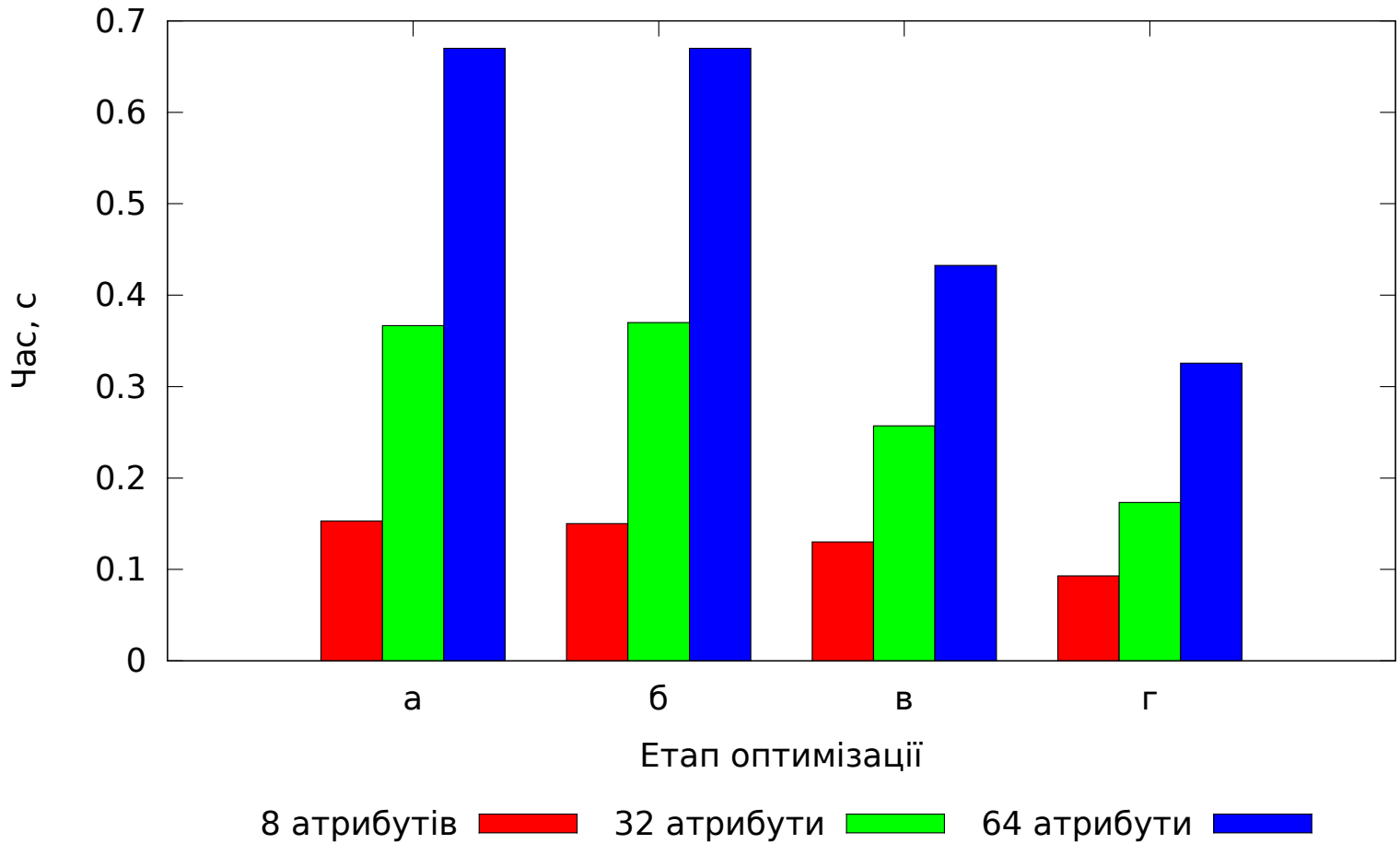
Neighbor-joining — оптимізації



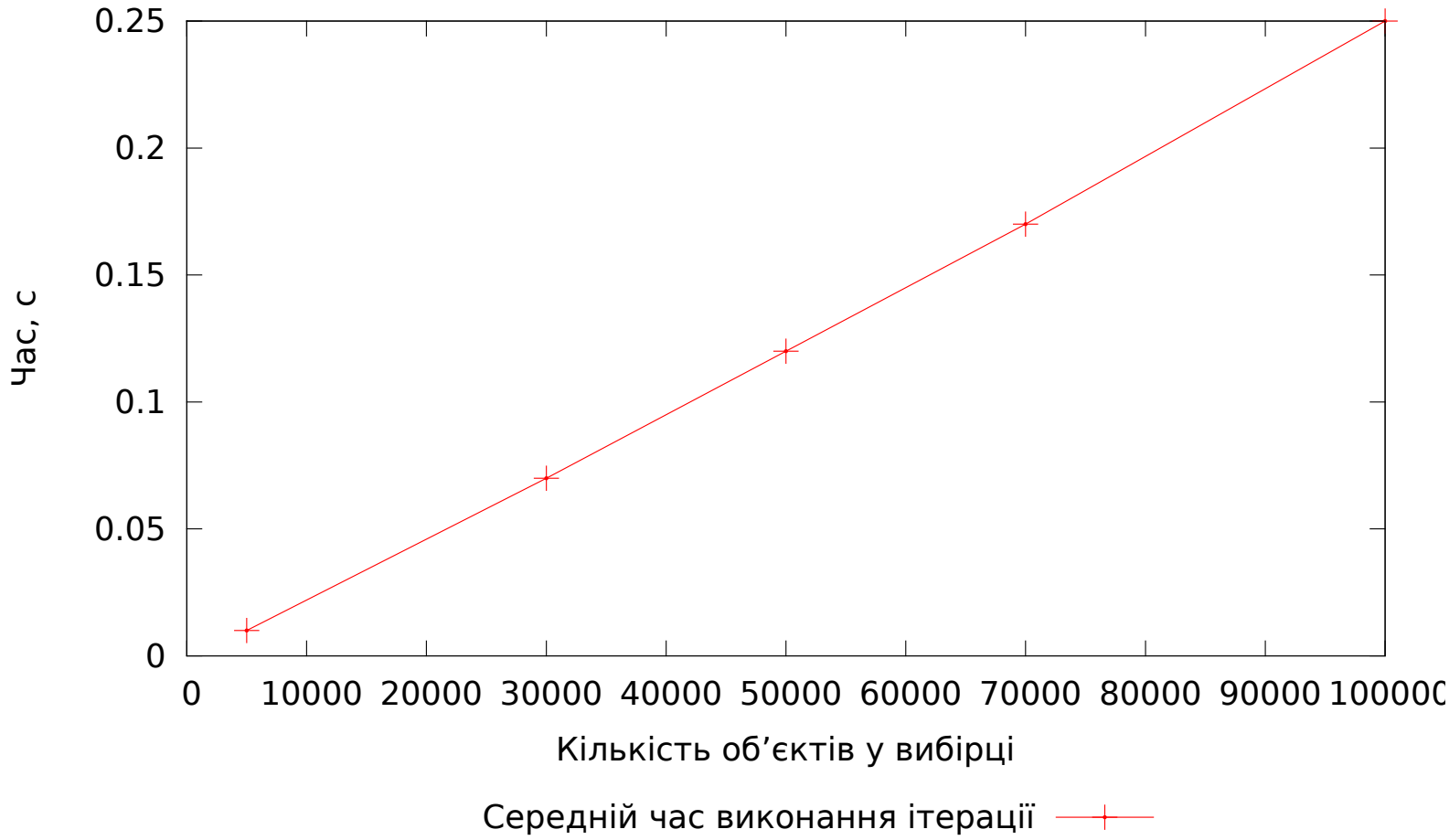
Neighbor-joining — залежність часу одної ітерації від розміру вхідних даних



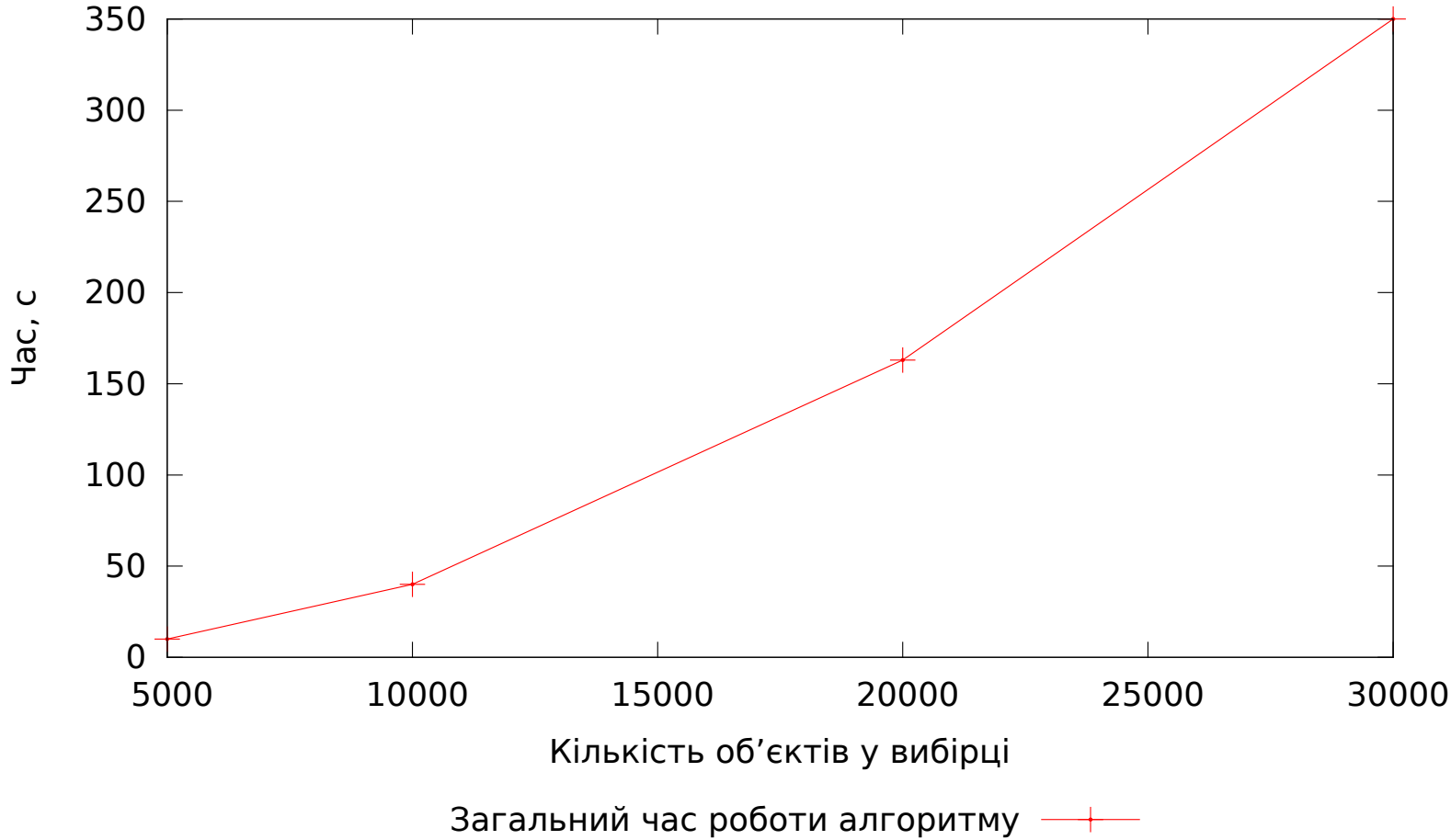
Neighbor-joining — оптимізації



Neighbor-joining — залежність часу одної ітерації від розміру вхідних даних



UPGMA — час ітерації



Висновки

Створено програмну реалізацію вищенаведених алгоритмів кластеризації та проведено оцінку їх ефективності для даних великих об'ємів.
Реалізація k-means здатна здійснити кластеризацію даних великих об'ємів із невеликими затратами часу. Обчислювальна складність задачі майже лінійно залежить від розміру вхідних даних.
Затрати часу на здійснення кластеризації за алгоритмом DBSCAN є значно більшими порівняно із k-means, але DBSCAN дозволяє отримати кластеризацію вищої якості.
Neighbor-joining дозволяє здійснити ієрархічну кластеризацію за прийнятний час.
UPGMA потребує надмірних затрат ресурсів.

Дякую за увагу!