Магістерська робота
Алгоритми кластеризації даних великих об'ємів

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

16 червня 2011 р.

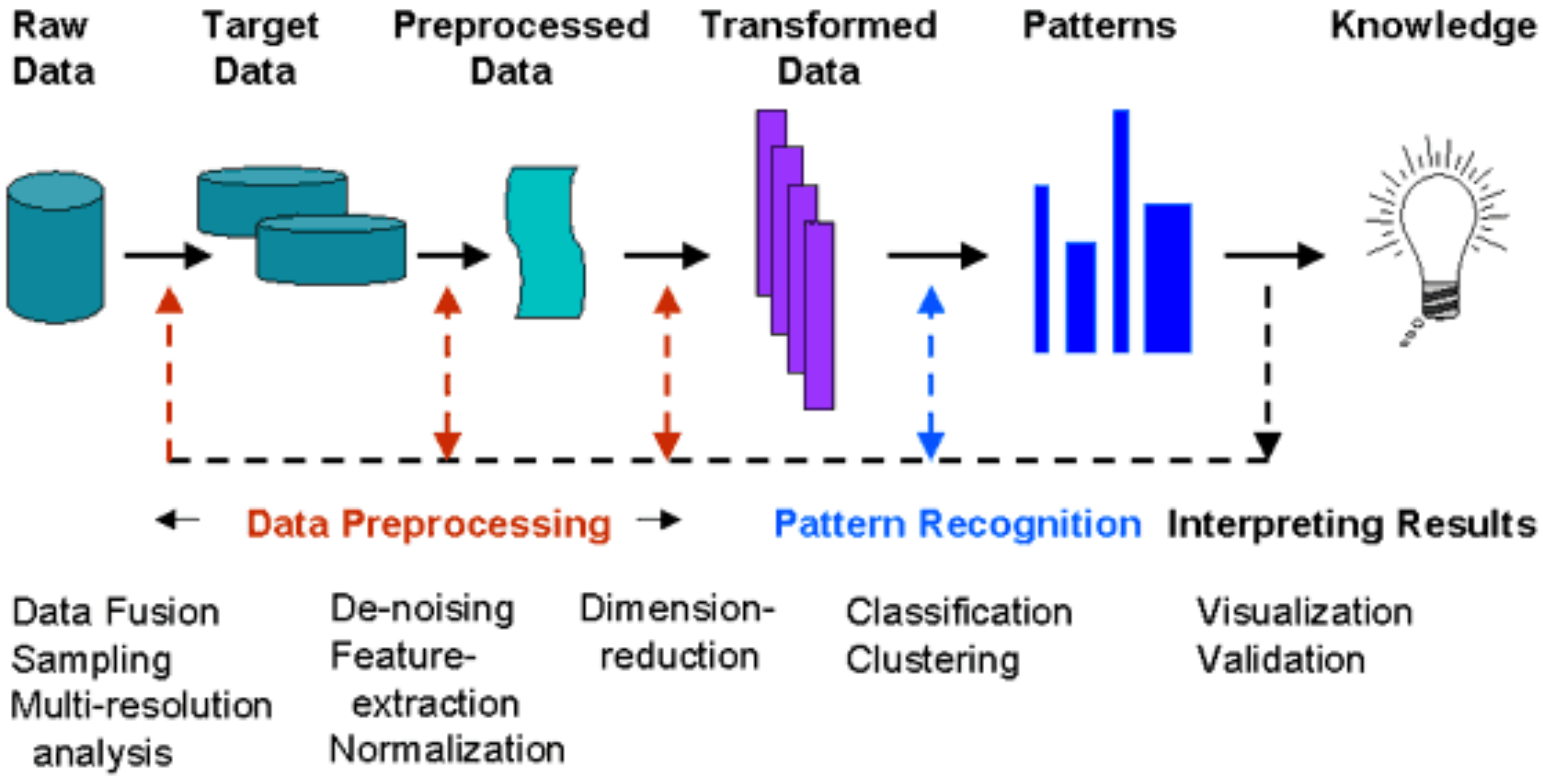
Задача кластеризації

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

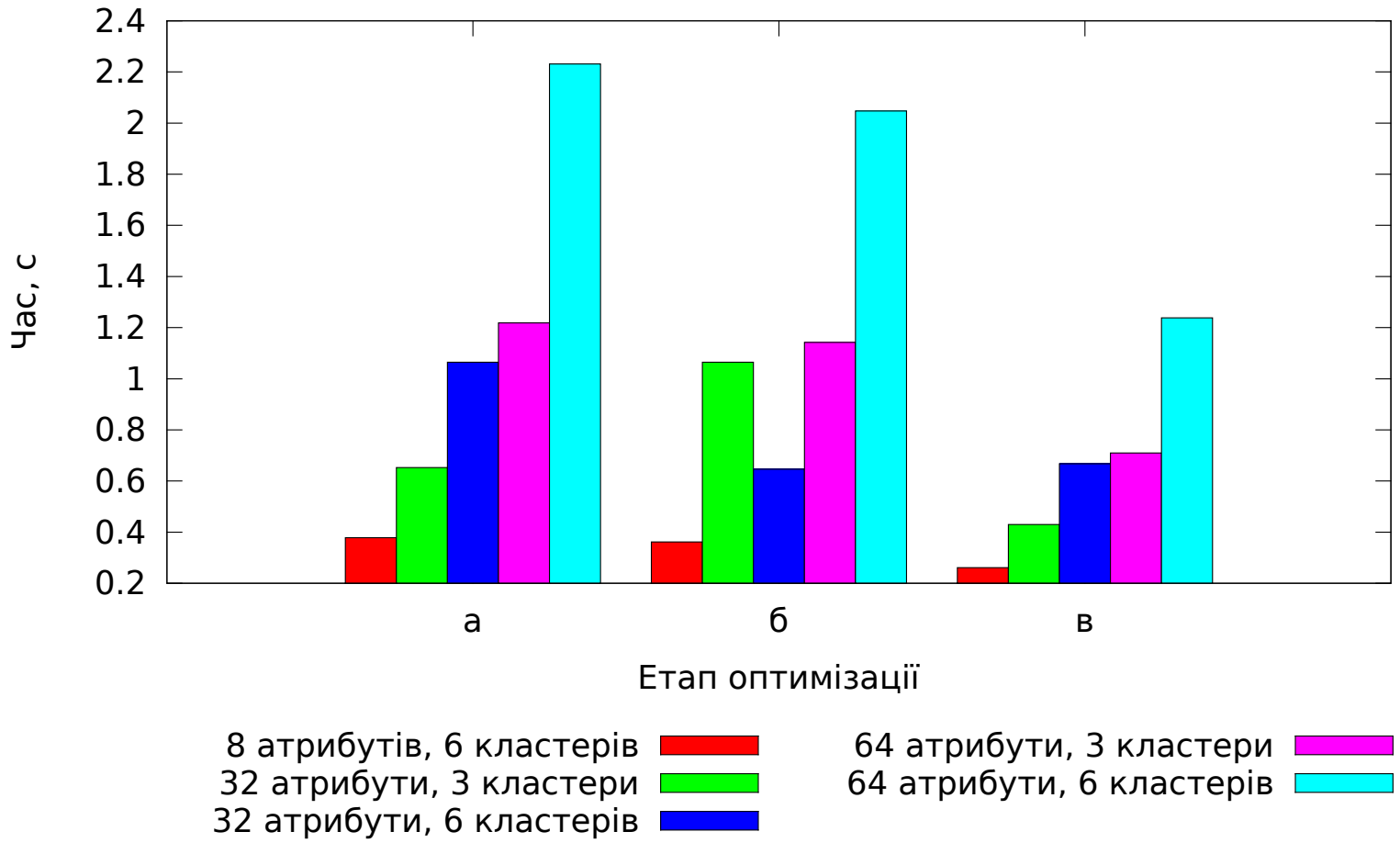
Тестові дані

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.

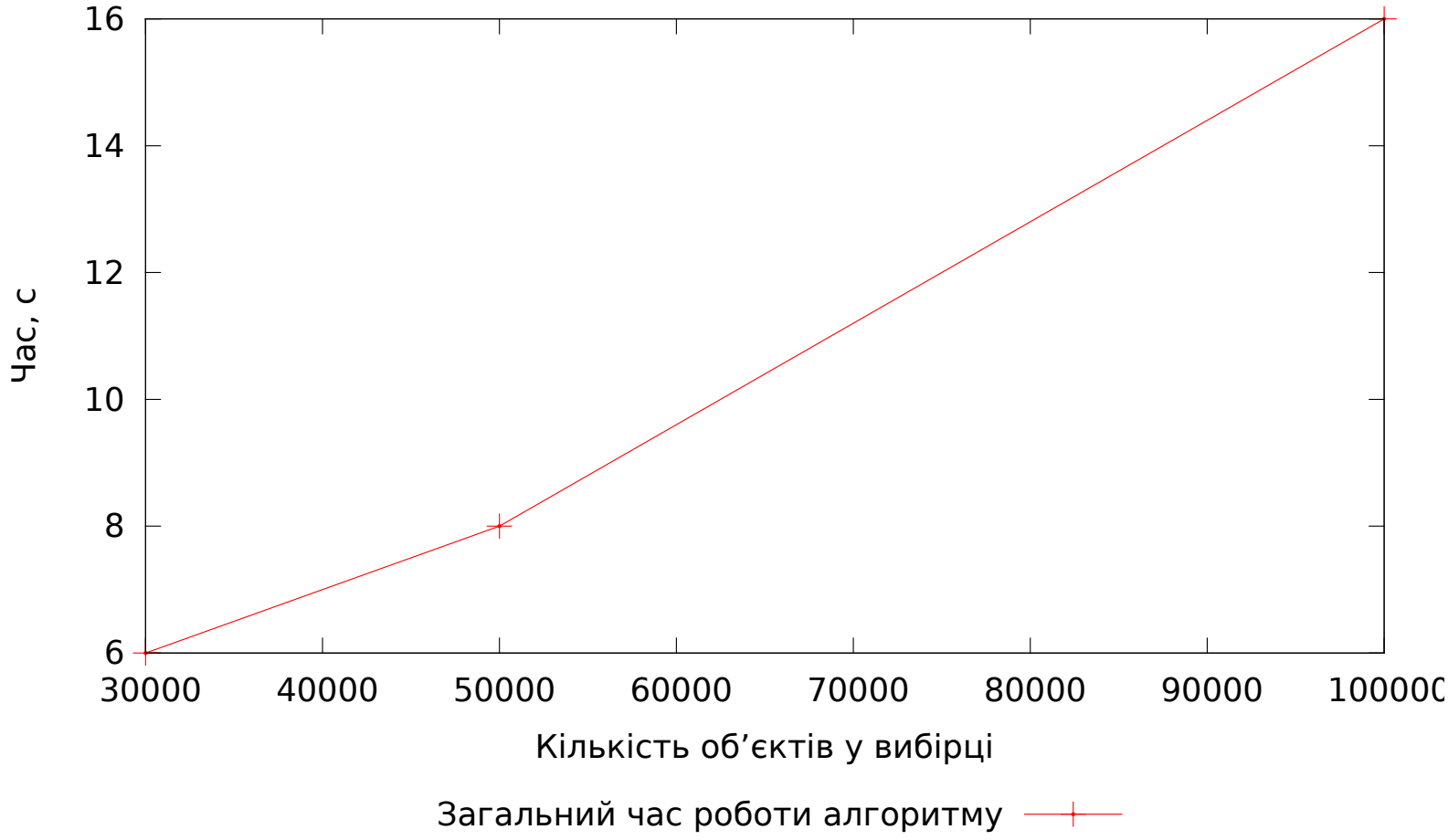
Критерії оцінки ефективності

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.

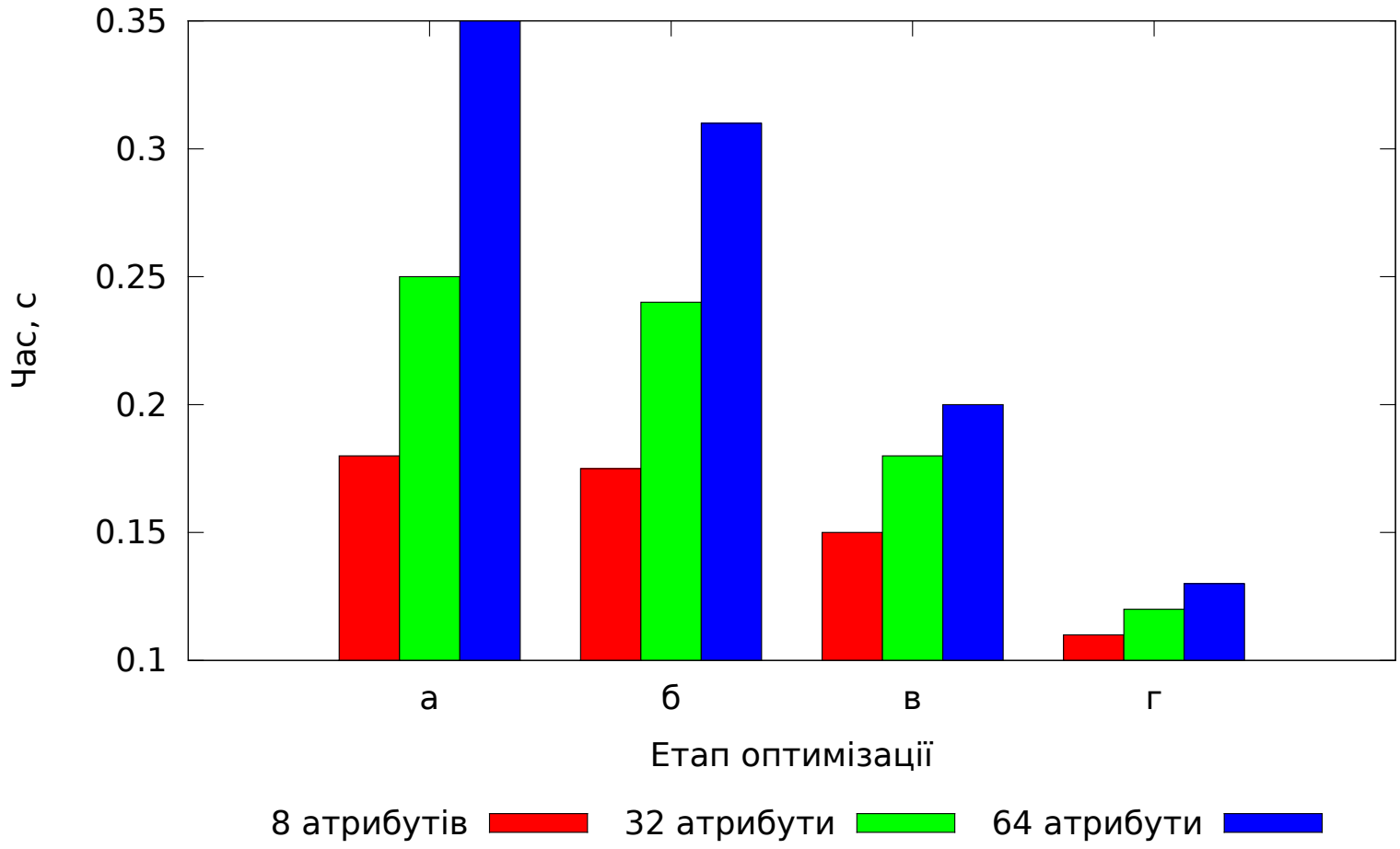
K-means — оптимізації



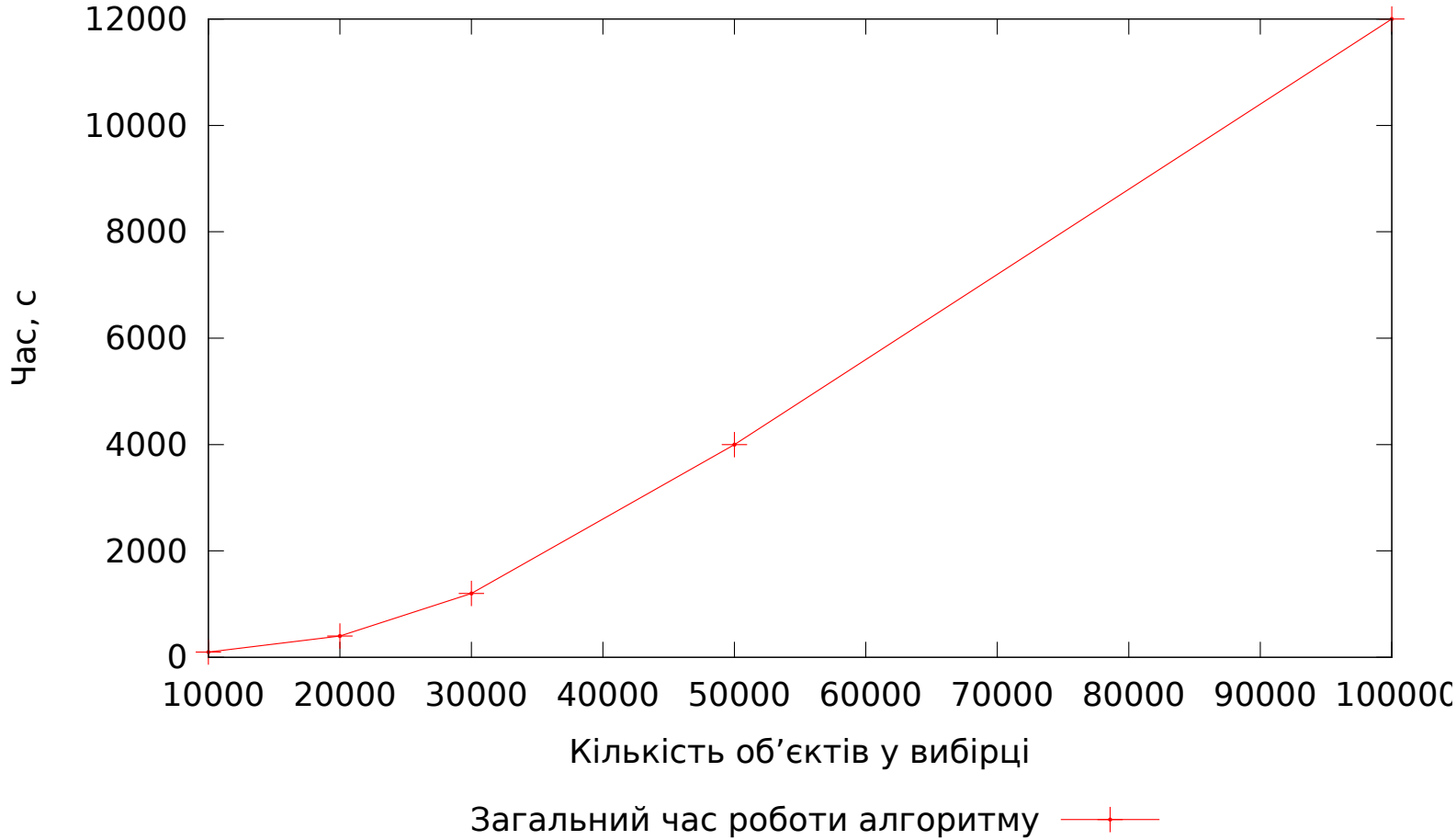
K-means — загальний час роботи



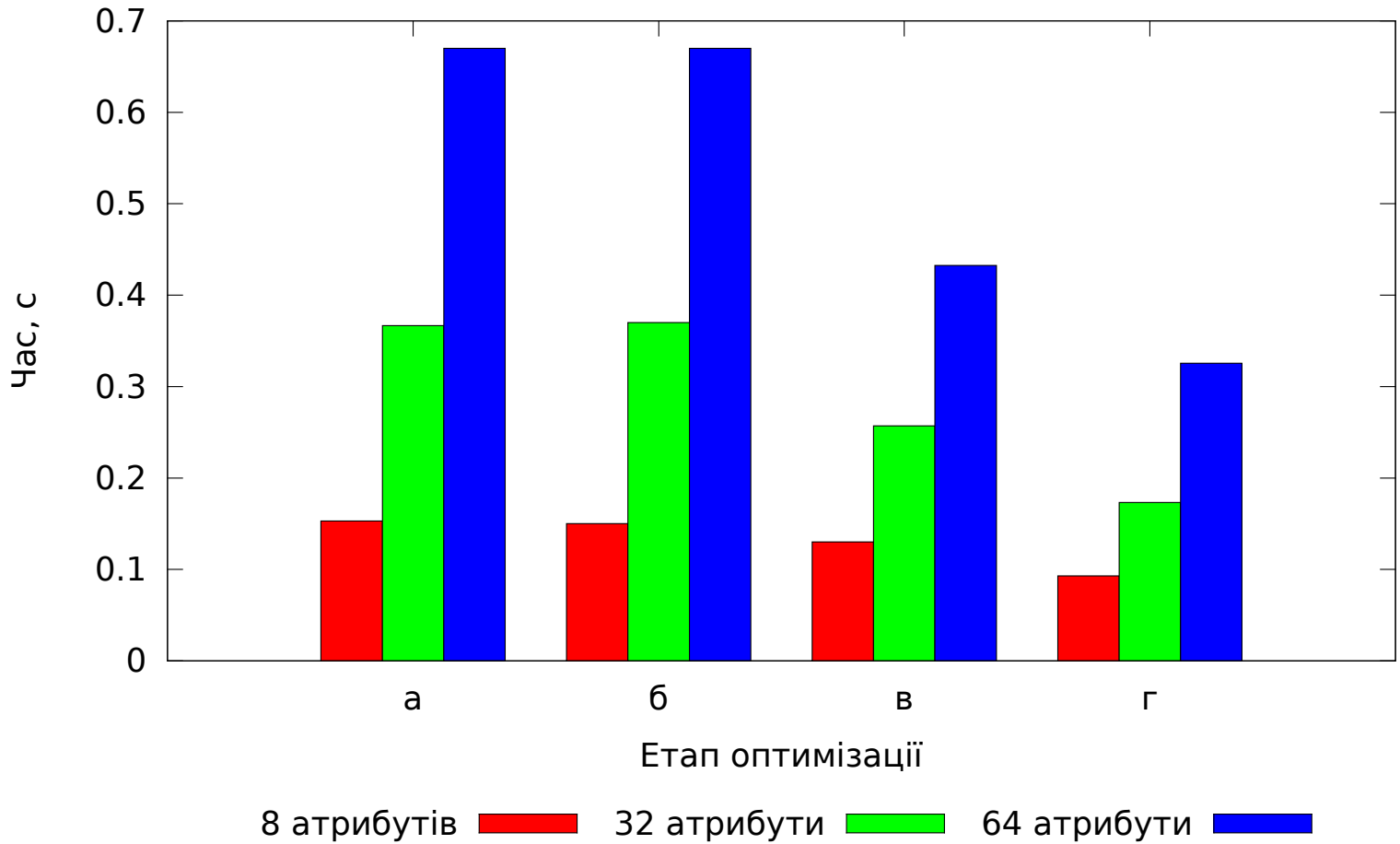
DBSCAN — оптимізації



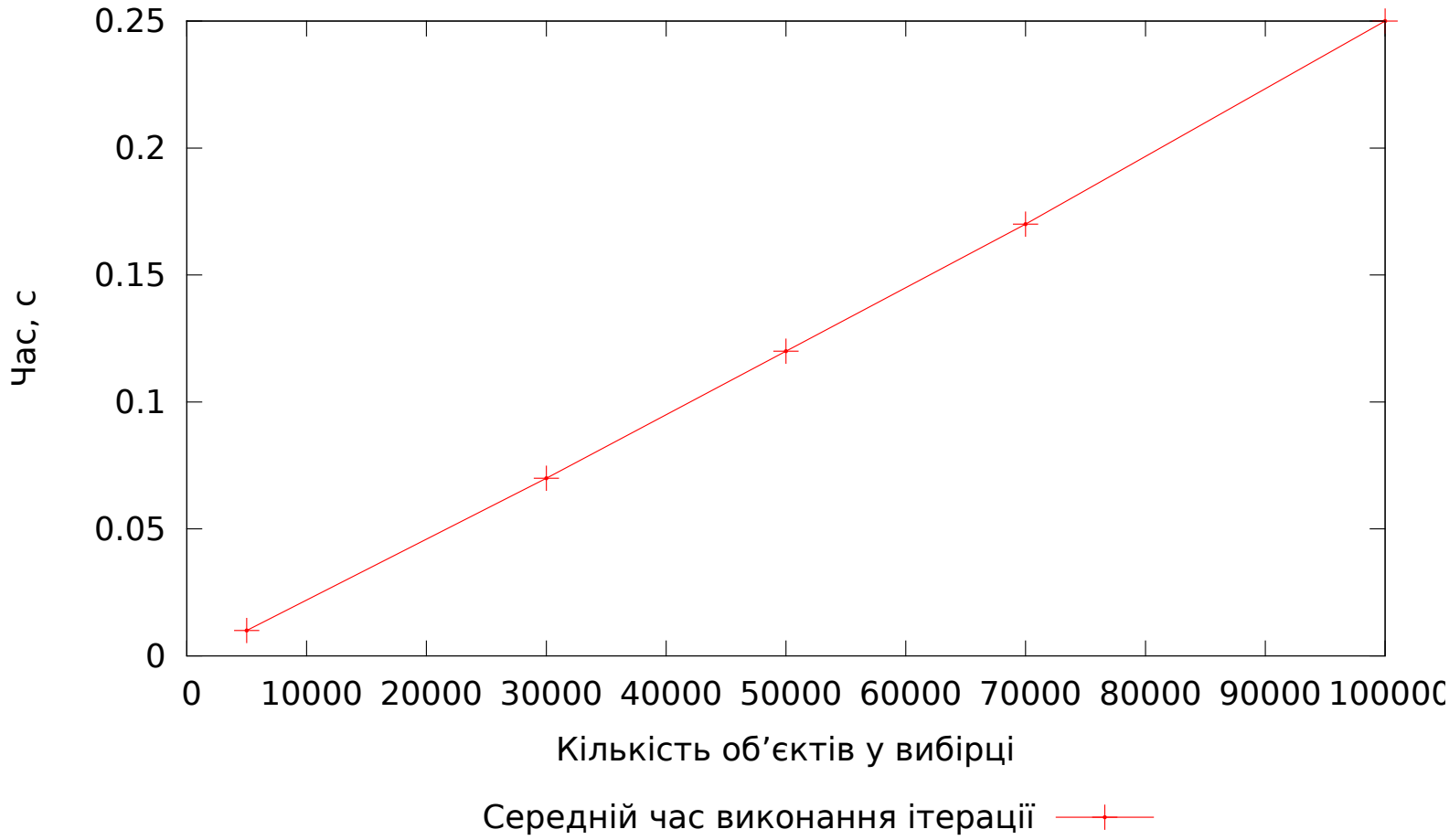
DBSCAN — загальний час роботи



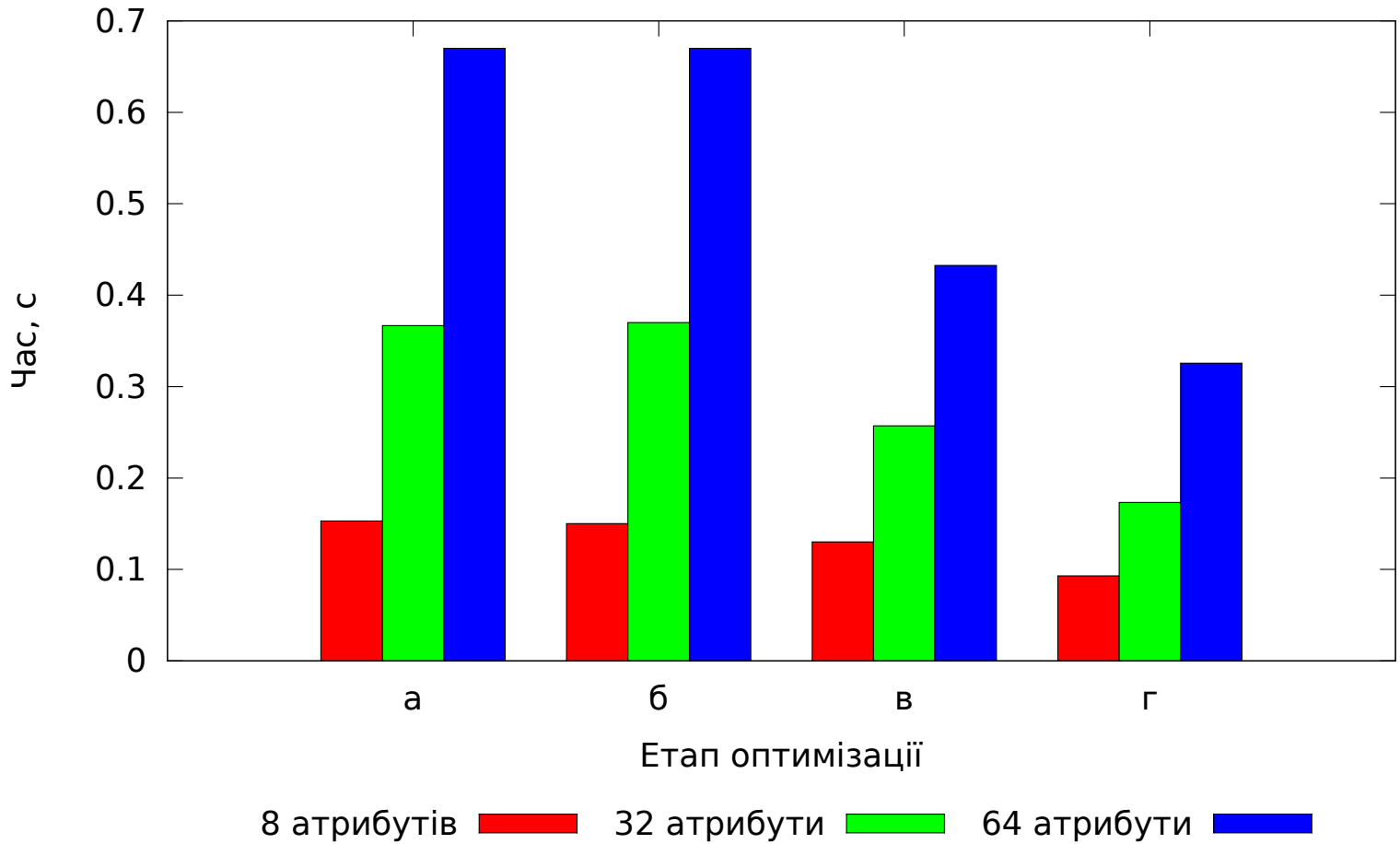
Neighbor-joining — оптимізації



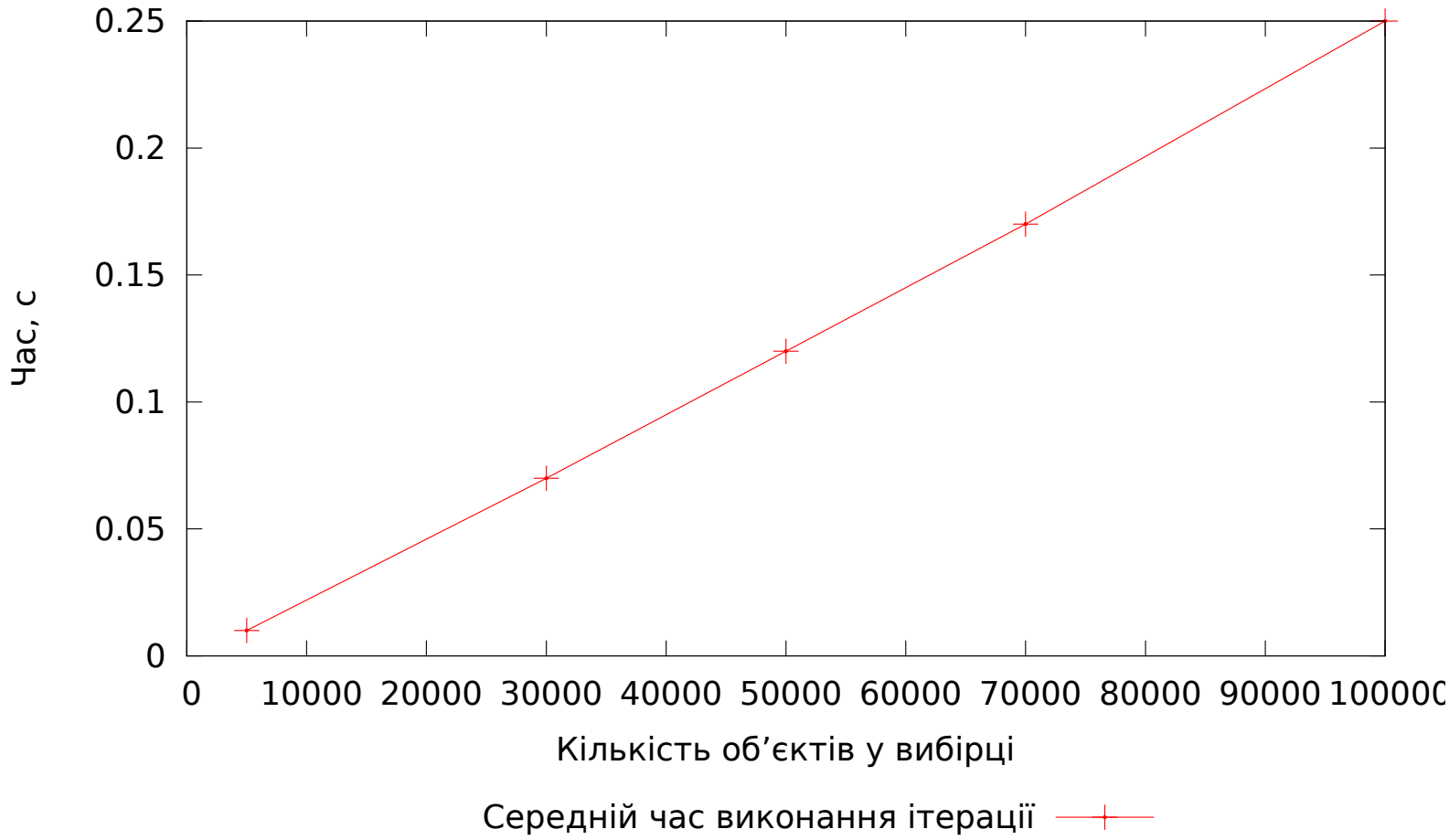
Neighbor-joining — залежність часу одної ітерації від розміру вхідних даних



Neighbor-joining — оптимізації



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



UPGMA — час ітерації