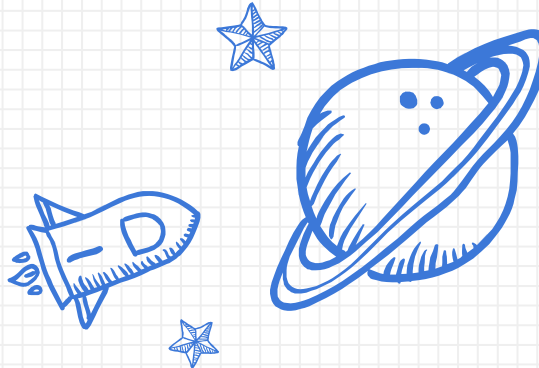


Large-Scale High-Dimensional Clustering with Fast Sketching

Antoine Chatalic, Rémi Gribonval, Nicolas
Keriven

Trần Khắc Việt
Hoàng





Giới thiệu

Giới thiệu bài toán

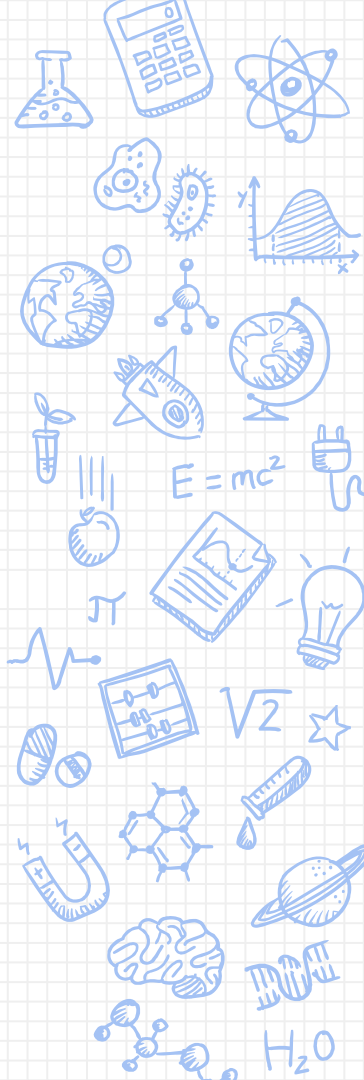
Vấn đề:

Giải quyết bài toán phân cụm k-means đa chiều với quy mô lớn.

Hướng giải quyết bài toán:

Với vấn đề “Quy mô lớn”: Sử dụng kỹ thuật phác thảo, nén toàn bộ tập dữ liệu vào một mô men tổng quát phi tuyến ngẫu nhiên.

Với vấn đề “đa chiều”: Sử dụng nhanh ma trận cấu trúc ngẫu nhiên để tính toán các toán tử phác thảo.



$$\text{Chc}\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$$

Xét việc phân cụm k-means bao gồm cả việc tìm k-

$$\mathcal{C} = \{c_1, \dots, c_k\} \subset \mathbb{R}^d$$

$$\text{SSE}(\mathcal{X}, \mathcal{C}) = \sum_{i=1}^n \min_j \|x_i - c_j\|^2. \quad (1) \quad \text{SE):}$$

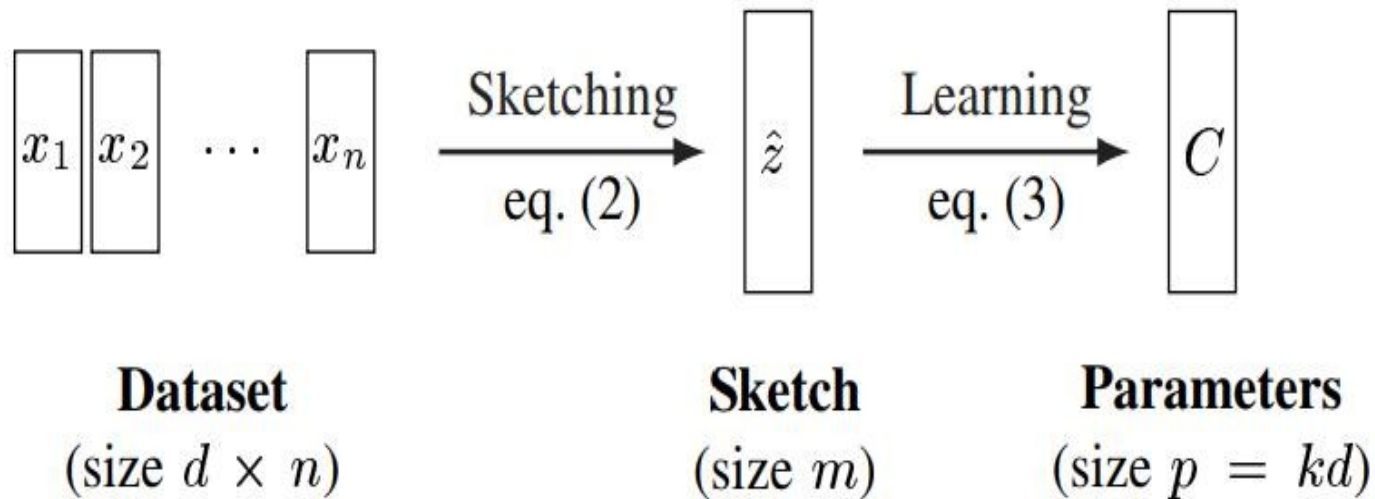
Một Framework đã được đề xuất để giải quyết các tập lớn bằng cách nén toàn bộ tập dữ liệu thành 1 véc tơ z

$$\hat{z} = \frac{1}{n} \sum_{i=1}^n \Phi(x_i), \text{ where } \Phi(x) = [e^{-i\omega_1^T x}, \dots, e^{-i\omega_m^T x}]^T. \quad (2)$$

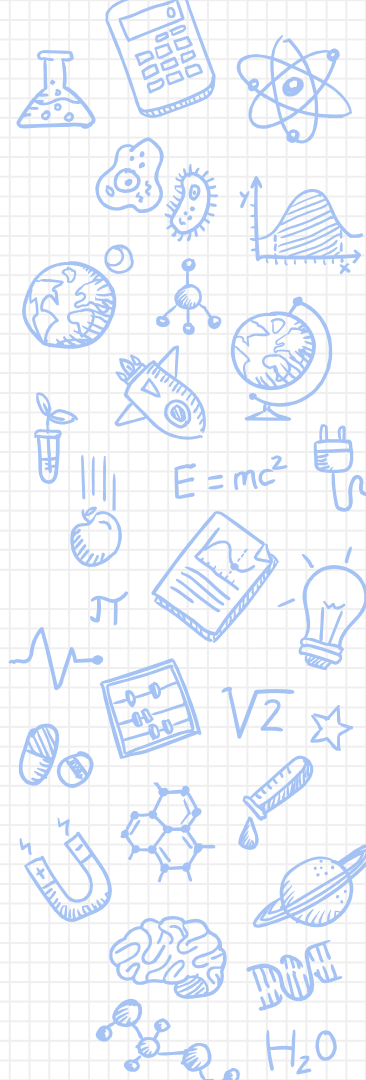
Từ đó centroid S và các giá trị liên quan có thể được tính toán hiệu quả:

$$\mathcal{C}, \alpha \in \arg \min_{\mathcal{C}, \alpha} \left\| \hat{z} - \sum_{i=1}^k \alpha_i \Phi(c_i) \right\|_2. \quad (3)$$

Mô tả bài toán



Tổng quan về quy trình công việc chung.





Áp dụng phép biến đổi nhanh

- ✗ Kết quả từ quá trình phân thảo có thể được đưa vào bằng cách tính toán các ma trận sản phẩm $W^T X$.
- ✗ Chứng minh lợi ích của việc thay thế ma trận tần số dày đặc W bằng ma trận cấu trúc ngẫu nhiên W_f .
- ✗ Sử dụng công thức $W_f = G_f R_f$
với R_f tương tự như R để tái chuẩn hóa và
 G_f là một thay thế nhanh của ma trận
Gaussian.

- 3

- ✗ Trường hợp tổng quát: xây dựng ma trận $m \times d$ tùy ý với $m > d$.

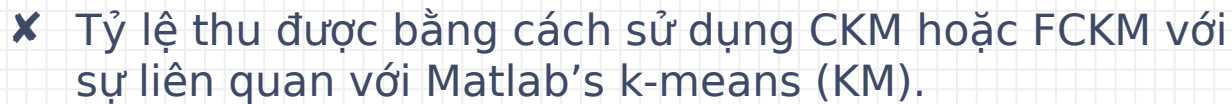
$$\lceil \log_2(d) \rceil \quad \lceil m/2^q \rceil$$
- ✗ Biểu thị $q = \lceil \log_2(d) \rceil$, $r = \lceil m/2^q \rceil$, $d_{pad} = 2^q$ và $m_{pad} = r2^q$.
- ✗ Phác thảo ma trận $W_f d_{pad} \times m_{pad}$ mà ma trận chuyển vị W_f^T được xây dựng bằng cách xếp chồng theo chiều dọc các khối vuông có kích thước $2^q \times 2^q$.
- ✗ Chi phí lưu trữ là $4rd_{pad} = 4m_{pad}$.

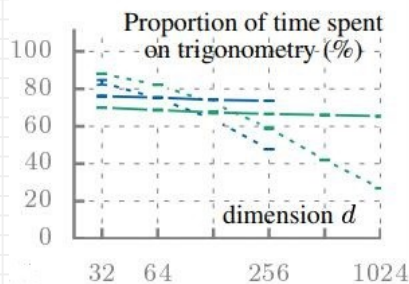
- ✗ Một phương pháp phân cấp đã được giới thiệu cho học hỗn hợp Gaussian.
- ✗ Bao gồm trong việc học hỗn hợp của Gaussians với hiệp phương sai đường chéo bằng cách đệ quy tách từng Gaussian theo hai chiều dọc theo chiều cao của phương sai cao nhất.
- ✗ Sau đó chỉ cần sử dụng các trung tâm của Gaussians khởi tạo để giảm thiểu các chức năng mất mát

Các thử nghiệm

- ✗ Thực hiện phép biến đổi nhanh với bộ công cụ SketchIbox Matlab.
- ✗ Sử dụng phép biến đổi Fast Walsh-Hadamard thích ứng của dự án Spiral.
- ✗ Trong quá trình thực nghiệm, chạy chương trình Matlab của k-means với chỉ số $l_{max} = 1000$.

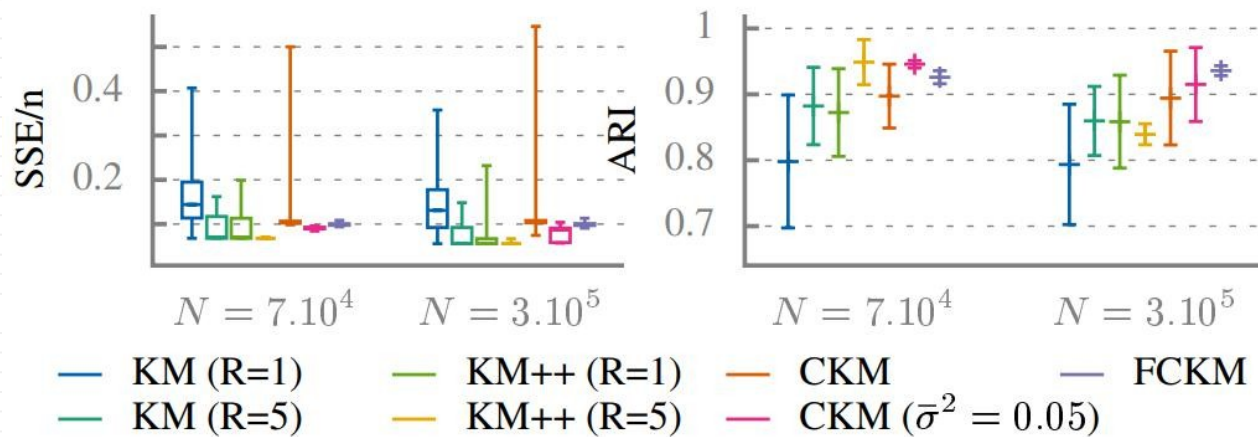
- ✗ Sử dụng dữ liệu nhân tạo.
- ✗ Thực hiện phân cụm trên $n = 10000$ véc tơ dữ liệu chạy ngẫu nhiên theo một hỗn hợp Gaussian với $k = 10$ với ma trận hiệp phương sai xác định.
- ✗ Chất lượng phân cụm được đo bằng chỉ số SSE.





- ✗ Thực hiện phân cụm trên dữ liệu MNIST của các chữ viết tay, trong đó có $k = 10$ lớp và kích thước $n = 7 \times 10^4$ ảnh.
- ✗ Các biến thể bị bóp méo được tạo ra bằng infMNIST là một tập có kích thước $n = 3 \times 10^5$.
- ✗ Với mỗi hình ảnh trích xuất mô tả SIFT được nối với một véc tơ đơn.
- ✗ Tính toán các ma trận tương tự giữa các véc tơ và k véc tơ đặc trưng đầu tiên của ma trận Laplacian để có được n phổ trong kích thước $d = k = 10$.

Phân cụm trên MNIST



✗ Kết quả giá trị của SSE và chỉ số điều chỉnh RAND (ARI) cho KM, CKM và FCKM.



~~✖~~ Tính toán các tính năng phổ biến qua hàm tích chập của phân hoạch riêng lẻ của kernel trên Laplacian thường rất tốn kém.

~~X~~ Tremblay et al. đề xuất bỏ qua bước này bằng cách sử dụng các tính năng $\log(k)$.

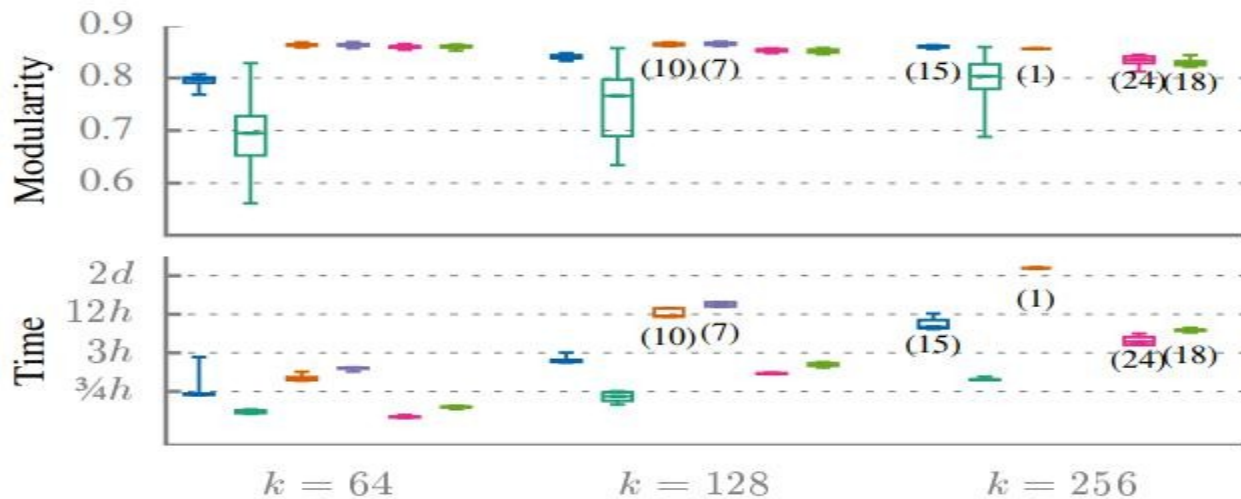
~~X~~ KM tiêu chuẩn sau đó được áp dụng trên một tập con ngẫu nhiên của các tính năng này và được nội suy

Xếp vào toàn bộ tập dữ liệu nhiên này với FCKM framework.

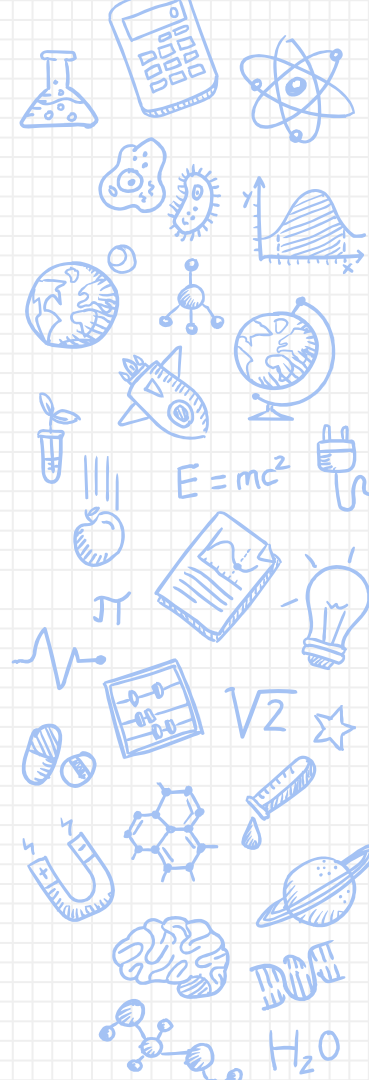
- ✗ Kết hợp nhanh các tính năng ngẫu nhiên này với FCKM framework.

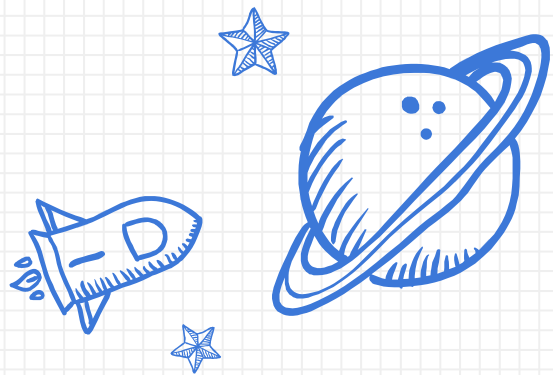
- ✗ Xét cộng đồng mua sắm của Amazon.
- ✗ Biểu đồ gồm $n = 334863$ nút và $E = 925872$ cạnh.
- ✗ Do không có công cụ dự đoán cho tập này nên lấy $k = 64, 128, 256$.
- ✗ So sánh cụm phổ ban đầu (SC), cụm phổ nén (CSC).
- ✗ Kết hợp 2 loại ma trận dày đặc và có cấu trúc với 2 quy trình học tập CL-OMPR và phân cấp (Hierarchical).

Phân cụm trên biểu đồ đồng mua



	Method	Features	Subs.	Sk. matrix	Clustering
—	SC	spectral	No	n/a	KM
—	CSC	random	Yes	n/a	KM
—	S2C	random	No	Dense	CL-OMPR
—	FS2C	random	No	Structured	CL-OMPR
—	HS2C	random	No	Dense	Hierarchical
—	HFS2C	random	No	Structured	Hierarchical



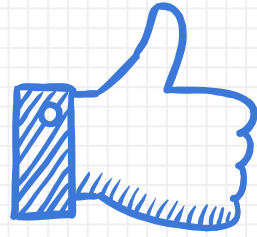


Triển vọng

X Để xuất một tổ chức để kết hợp hiệu quả mà trên có cấu trúc phân cấp với khả năng tiếp nhận của các phương pháp tiếp cận dựa trên phân tích và các phương pháp xử lý nghiệp vụ của phân cấp, tạo ra một framework có thể xử lý các tập lớn và đa chiều với giới hạn mức chi phí dung lượng thấp. **X** Machine learning là một hàm Kernel.

✖ Nhưng trên thực tế sử dụng để phân tích nhiễu ngẫu nhiên là tìm một quy tắc hàm Kernel số lượng nhiều các centroid hơn là CL-OMPR tuy nhiên k vẫn là một hàm bậc 2 khi $m = \theta(kd)$.

- ✗ Khó khăn : dù thuật toán phân cấp đã được đưa ra để cho phép giải quyết bài toán với số lượng nhiều các centroid hơn là CL-OMPR tuy nhiên k vẫn là một hàm bậc 2 khi $m = (kd)$.



**THANK YOU FOR
YOUR ATTENTION !**



Thảo luận và Q/A.