K-MEANS TRÊN MAPREDUCE

Đặng Quang Anh - Nguyễn Trung Hiếu - Võ Huy Quang - Lê Thuý Quỳnh - Hoàng Minh Tuấn

Tháng 4 2021

Tóm tắt nội dung

K-Means là giải thuật phân cụm dữ liệu khá nổi tiếng và được sử dụng phổ biến trong lĩnh vực khai phá dữ liệu, nó cho phép chia n đối tượng thành k cụm sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm là nhỏ nhất. Tuy nhiên, phương pháp này còn nhiều hạn chế do việc tính khoảng cách giữa các đối tượng đến các tâm và xác định lại tâm được thực hiện lặp lại nhiều lần khiến giải thuật mất nhiều thời gian xử lí và khó triển khai trên tập dữ liệu lớn.
Vì vậy chúng ta cần đến mô hình tính toán song song MapReduce để triển khai K-Means trên tập dữ liệu lớn.

Outline

Thuật toán phân cụm K-Means

Mô hình MapReduce

Thuật toán MRK-MEANS (MAPREDUCE K-MEANS)

Thuật toán phân cụm K-Means

Thuật toán phân cụm K-Means

Mô hình MapReduce

Thuật toán MRK-MEANS (MAPREDUCE K-MEANS)

Thuật toán phân cụm K-Means

1. Khái niệm giải thuật K-Means:

- Là một thuật toán phân cụm đơn giản được sử dụng để giải quyết bài toán phân cụm
- Giải thuật gom cụm K-Means được sử dụng để phân loại tập dữ liệu phi cấu trúc hoặc bán cấu trúc, là 1 trong những phương thức phân loại dữ liệu phổ dụng và hiệu quả do tính đơn giản và khả năng kiểm soát tập dữ liệu

Thuật toán phân cụm K-Means

2. Ứng dụng giải thuật K-Means:

- Thuật toán K-Means thường được sử dụng trong các ứng dụng:
  - Thống kê dữ liệu
  - Công cụ tìm kiếm(Search engine)
  - Phân loại khách hàng

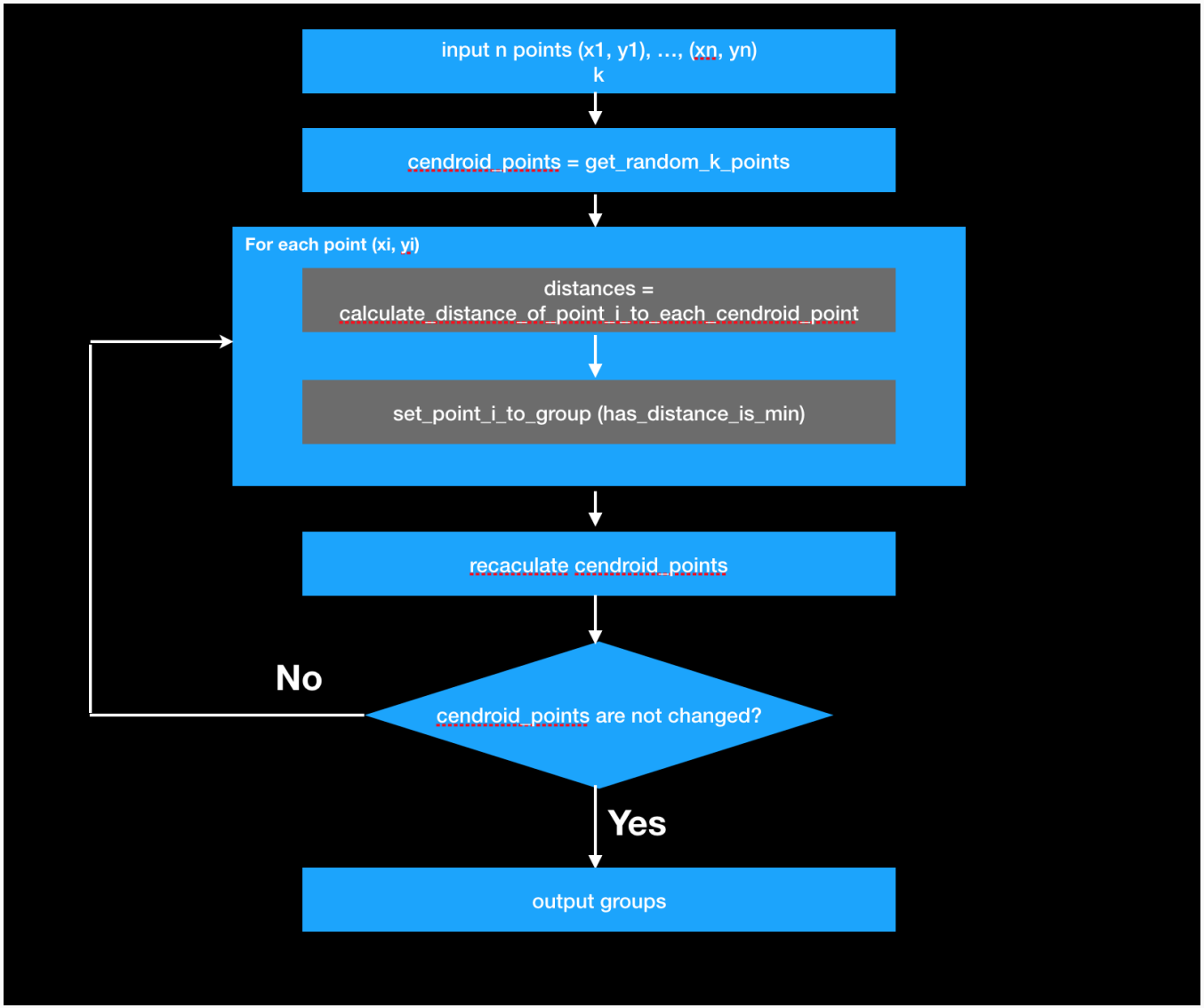
Thuật toán phân cụm K-Means

3. Phương pháp giải thuật K-Means

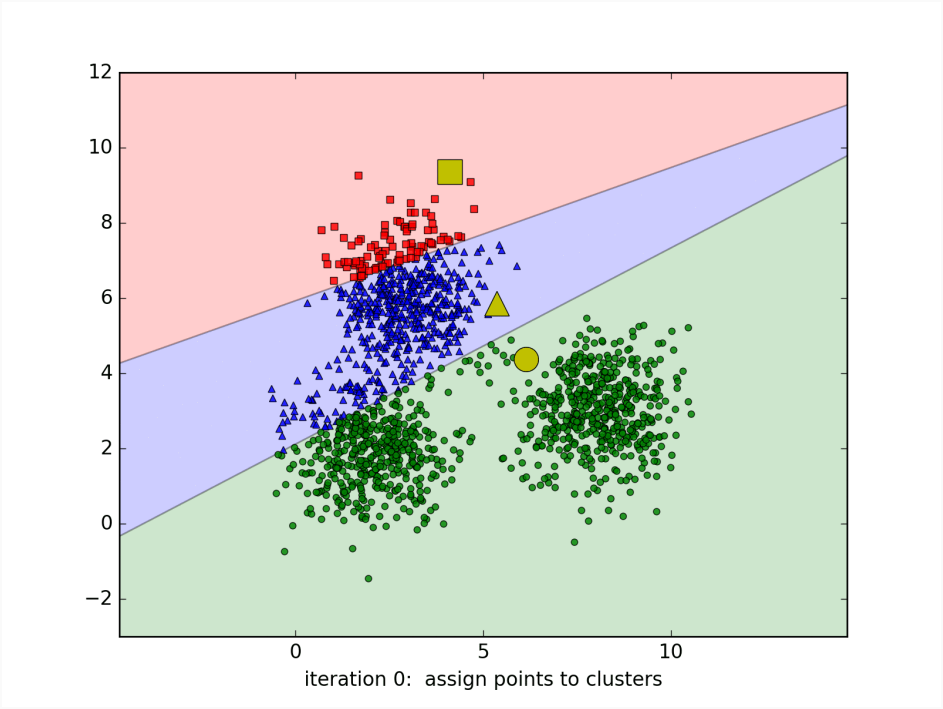
- **Bước 1:** Chọn k điểm bất kì trong tập dữ liệu làm tâm ban đầu.
- **Bước 2:** Với các điểm dữ liệu còn lại được gán cho cụm có khoảng cách đến tâm gần nhất.
- **Bước 3:** Tính toán lại khoảng cách để cho tâm cụm gần chính giữa cụm nhất
- Lặp lại bước 2 và 3 đến khi vị trí các tâm không đổi.

$=>$ **Nhược điểm:** Tiêu tốn tài nguyên hệ thống và thời gian, chỉ phù hợp với dữ liệu nhỏ và vừa.

Thuật toán phân cụm K-Means



Thuật toán phân cụm K-Means



Mô hình MapReduce

Thuật toán phân cụm K-Means

Mô hình MapReduce

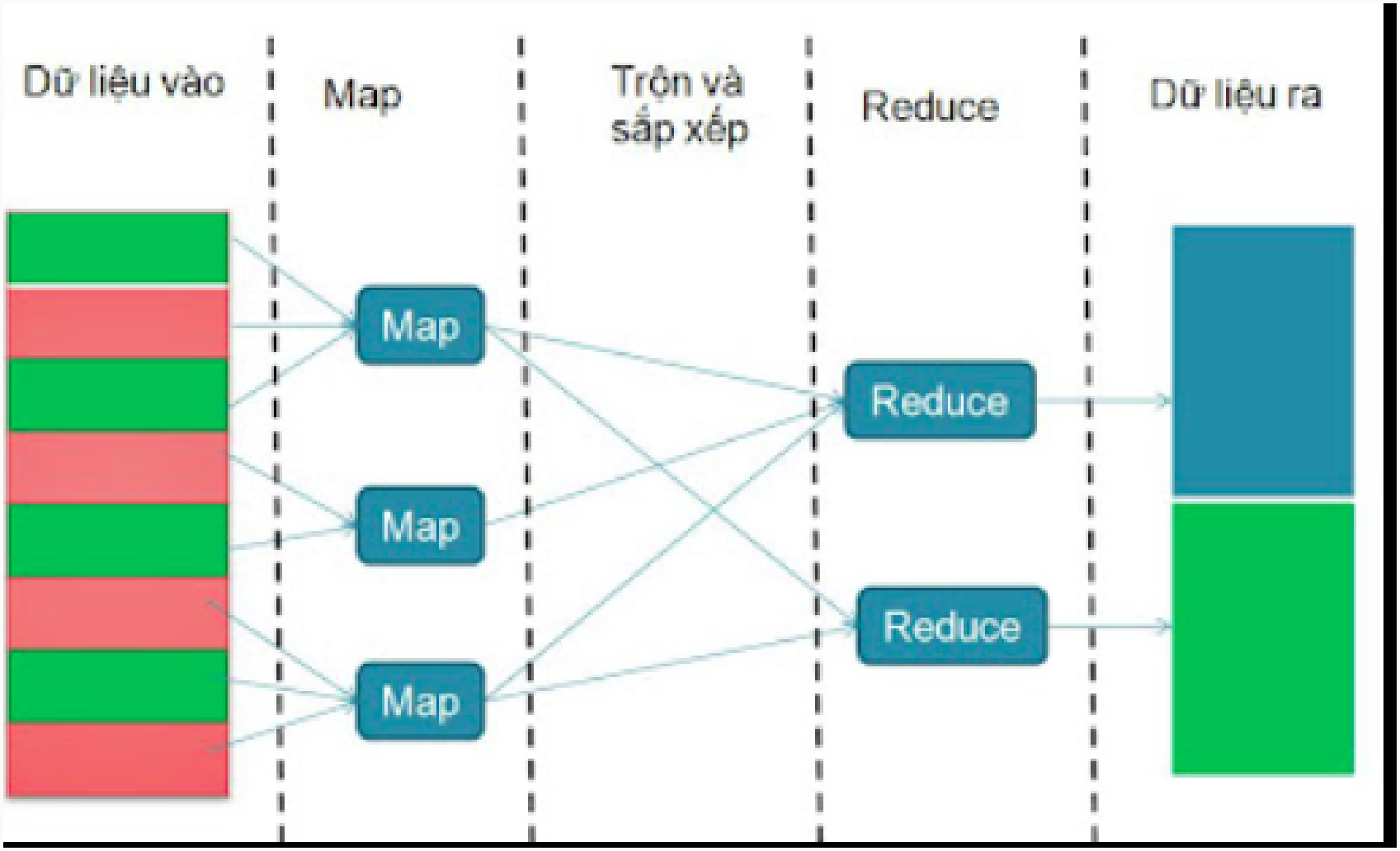
Thuật toán MRK-MEANS (MAPREDUCE K-MEANS)

Mô hình MapReduce

1. Khái niệm mô hình MapReduce

- Là mô hình lập trình dùng để tính toán trên tập dữ liệu lớn
- Một trình xử lý MapReduce có thể tính toán đến terabytes hoặc petabytes dữ liệu trên hệ thống được kết nối thánh cụm các nodes.
- Ứng dụng:
  - Thống kê số lượt truy cập URI.
  - Thống kê số từ khoá trên các hotnames.
  - ...
- MapReduce gồm 2 pha: Map và Reduce.

Mô hình MapReduce



Mô hình MapReduce

2. Bộ ánh xạ (Mapper):

- Xử lý tập dữ liệu đầu vào dưới dạng (*key*, *value*) và tạo lập trung gian là cặp (*key*, *value*)
- Bộ ánh xạ thực hiện 3 bước sau :
  - Bước 1: Ánh xạ cho mỗi nhóm dữ liệu đầu vào dạng (*key*, *value*)
  - Bước 2: Xử lý đầu vào -> chia nhóm -> tạo ra các (*key*, *value*) trung gian.
  - Bước 3: Đầu ra bộ ánh xạ được lưu trữ và định vị trong mỗi bộ Reducer.

Mô hình MapReduce

3. Bộ Reducer:

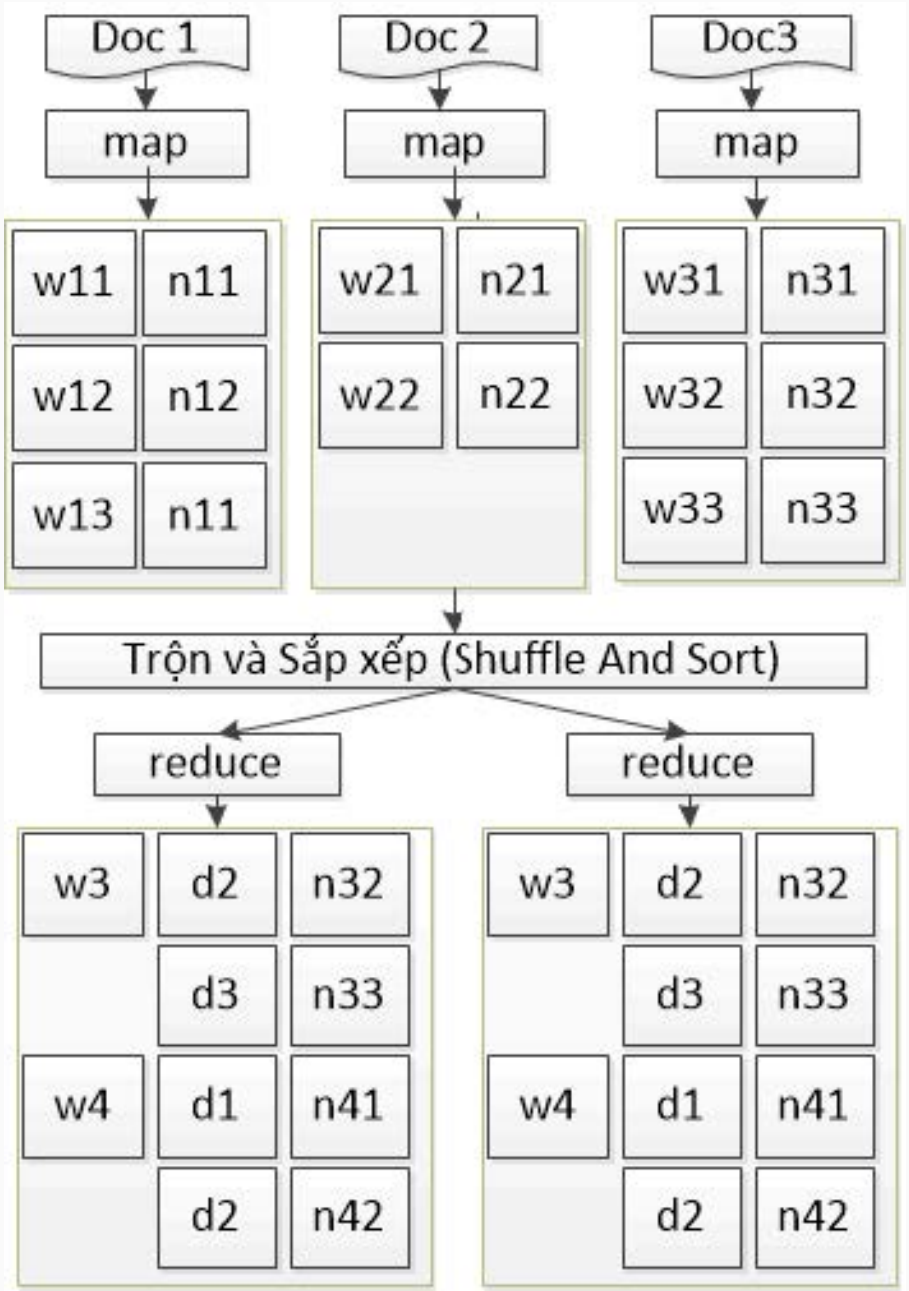
- Trộn tất cả các phần tử có cùng *key* trong tập dữ liệu tạo ra bởi Mapper để hình thành tập giá trị nhỏ hơn.
- Bộ Reducer thực hiện 3 bước sau :
  - Bước 1: Đầu vào của Reducer là đầu ra của Mapper. MapReduce sẽ gắn khối liên quan cho bộ Reducer.
  - Bước 2: Đầu vào của Reducer được gom theo *key* -> giai đoạn trộn và sắp xếp.
  - Bước 3: Tạo bộ so sánh để nhóm các *key* trung gian lần 2 nếu quy luật gom nhóm khác với trước đó.

Mô hình MapReduce

4. Hoạt động của MapReduce

- Đọc dữ liệu đầu vào.
- Xử lý dữ liệu đầu vào (Map).
- Sắp xếp và trộn các kết quả thu được từ các máy tính phân tán thích hợp.
- Tổng hợp các kết quả trung gian thu được (Reduce).
- Đưa ra kết quả cuối cùng.

Mô hình MapReduce



Thuật toán MRK-MEANS (MAPREDUCE K-MEANS)

Thuật toán phân cụm K-Means

Mô hình MapReduce

Thuật toán MRK-MEANS (MAPREDUCE K-MEANS)

Cách hoạt động của thuật toán MRK-MEANS

Thuật toán K-Means dựa trên mô hình MapReduce như sau:

Bước 1: Partition:

Khởi tạo k tâm ban đầu và chia tập dữ liệu thành các tập dữ liệu con nhỏ hơn

Bước 2: Local clustering:

Thực hiện tính toán để phân cụm trong từng bộ dữ liệu con

- Mapper: Đọc dữ liệu x_i và tìm tâm y_j gần với x_i nhất. Tức là, $j = argmin_i ||x_i - y_j||_2$ và phát ra $< j, x_i >$
- Combiner: Lấy $< j, \{x_i\} >$ và phát ra $< j, \Sigma x_i, num >$ trong đó num là số đối tượng mà Combiner nhận khóa j

Cách hoạt động của thuật toán MRK-MEANS

Thuật toán K-Means dựa trên mô hình MapReduce như sau:

Bước 3: Global clustering:

Gộp cụm ở bộ dữ liệu lớn và tính toán lại tâm mới

- Reducer: Lấy $< j, \{(\Sigma x_i, num)\} >$ và tính tâm mới, phát ra tâm mới $< j, y_j >$ của cụm