# TỐI ƯU DỮ LIỆU LỚN HÀNG HẢI GOM CỤM K NHÓM THEO TRUNG BÌNH DỰA VÀO MÔ HÌNH MAPREUCE

# OPTIMIZED THE MARITIME BIG DATA K-MEANS CLUSTERING BASED ON THE MAPREDUCE MODEL

**PHẠM TUẤN ANH[1,2], ĐẶNG XUÂN KIÊN[1], PHẠM TÂM THÀNH[3,*]**

*[1]Trường Đại học Giao thông vận tải Thành phố Hồ Chí Minh*

*[2]Tổng Công ty Bảo đảm an toàn Hàng hải Miền Nam*

*[3]Trường Đại học Hàng hải Việt Nam*

**Email liên hệ: phamtamthanh@vimaru.edu.vn*

**Tóm tắt**

*Với sự phát triển của công nghệ thông tin, dữ liệu hàng hải lớn đang là xu hướng ngày càng tăng của các ứng dụng nhằm xử lý mà không đủ bộ nhớ chính của việc phân tích dữ liệu lớn đang là bài toán thách thức hiện nay. Đối với ứng dụng chuyên sâu, dữ liệu hàng hải lớn, thuật ngữ "MapReduce" gần đây đã thu hút sự chú ý đáng kể và bắt đầu được nghiên cứu để phân tích mà có thể xử lý hàng petabyte dữ liệu AIS cho hàng triệu tàu thuyền. MapReduce là một mô hình lập trình cho phép dễ dàng phát triển các ứng dụng song song có thể mở rộng để xử lý dữ liệu lớn trên các cụm máy tính [1]. Trong bài nghiên cứu này, một thuật toán gom cụm được gọi là K-means dựa trên mô hình MapReduce để xử lý dữ liệu hàng hải tàu biển tại khu vực miền Nam, Việt Nam. Với kết quả thu được, chúng tôi đưa ra suy luận hoặc dự đoán về dữ liệu gom cụm mà chúng được thu thập và sau đó là hiển thị dữ liệu của các hàng hải tàu biển, bao gồm quy mô, hướng và phân bố không gian.*

**Từ khóa**: *Mô hình MapReduce, K-means, dữ liệu AIS, khai phá dữ liệu.*

**Abstract**

*With the development of information technology, the maritime big data is an increasing trend of applications being expected to deal with big data that usually do not fit in the main memory of an analyzing big data is a challenging problem today. For such data intensive application, the maritime big data, the "MapReduce" framework has recently attracted considerable attention and started to be investigated for analysis which can handle petabyte of AIS data for millions of vessels. MapReduce is a programming model that allows easy development of scalable parallel applications to process big data on large clusters of commodity machines. This study, a standard clustering algorithm called K-means is based on the MapReduce model to be processed the marine traffic data in southern region, Viet Nam. According to the main results obtained, we concerned with making inference or prediction the clustering data which were collected and were shown the dashboard of maritime vessels traffic, including the scale, the trend of change and the spatial distribution situation.*

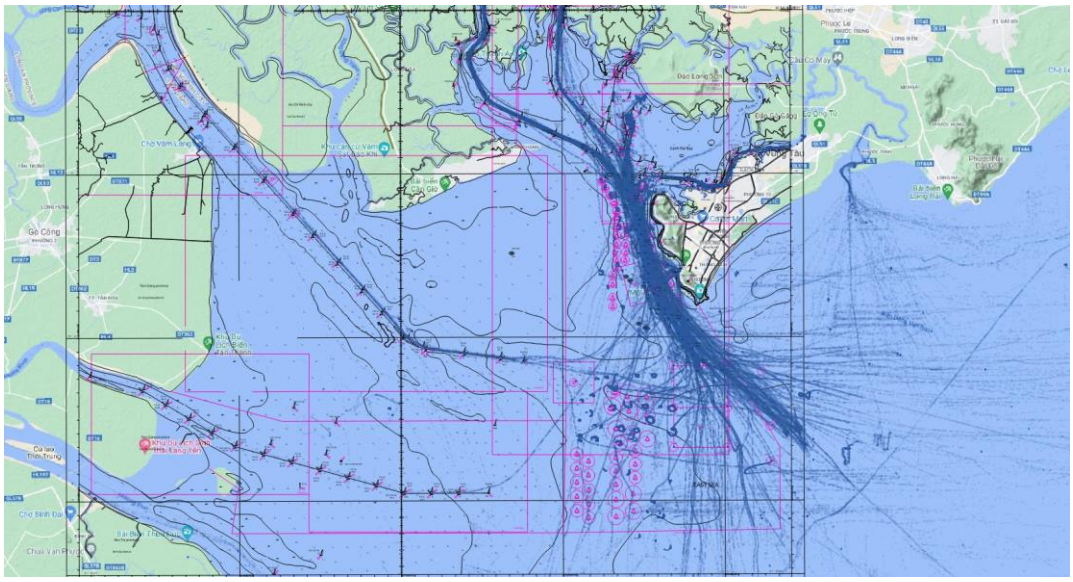
**Keywords**: *MapReduce, K-means, AIS data, data mining.*

## 1. Đặt vấn đề

Với sự phát triển mạnh mẽ của kinh tế biển, với mật độ tàu thuyền dày đặc, đặc biệt là tập trung các cảng biển lớn có khả năng tiếp nhận các tàu trọng tải lên tới 160,000DWT, điều này đã tạo ra dữ liệu lớn hàng hải [2]. Dữ liệu hàng hải được thu thập từ hệ thống thông tin nhận dạng tự động (AIS, Automatic Identification System) [3], cung cấp nhiều thông tin thời gian thực về hàng hải tàu biển và đã sử dụng để nhận thức tình huống hàng hải (MSA, Maritime Situation Awareness) và giám sát đại dương. Sự phổ biến của hệ thống AIS đồng nghĩa với cung cấp một nguồn dữ liệu phong phú để khai phá dữ liệu phục vụ phân tích giao thông hàng hải tàu biển, theo thống kê lượng dữ liệu được thu thập từ hệ thống AIS trong năm qua là rất lớn (tại khu vực miền Nam, Việt Nam đã thu thập hơn 100GB) - Trong nghiên cứu này, chúng tôi lấy mẫu dữ liệu ngày 13/9/2019. Được thể hiện trong Hình 1.

Dữ liệu hàng hải là dữ liệu được thu thập qua hệ thống AIS chứa nhiều thông tin tàu biển (thời gian, tên tàu, MMSI - Maritime Mobile Service Identity , COG-Course Over Ground, SOG - Speed Over Ground,...).

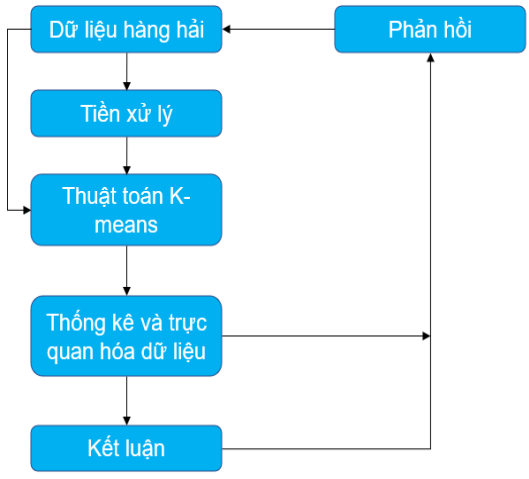
TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY



*Hình 1. Bản đồ dữ liệu hàng hải tàu biển trong ngày 13/9/2019 (có 6.310.956 thông báo AIS)*

Việc phân tích và nghiên cứu các dữ liệu lớn hàng hải này có thể tìm ra hành trình của tàu biển như vị trí, hành vi điều hướng tàu một cách nhanh chóng, tự động và thông minh. Qua đó, các tổ chức hàng hải định hướng cho sự phát triển hiệu quả hoạt động của ngành hàng hải, đóng góp vào sự phát triển kinh tế biển. Từ các dữ liệu được thu thập, trong đó một số dữ liệu được lặp lại cùng với dữ liệu lớn vị trí của tàu biển dẫn đến hai thách thức đối với việc sử dụng dữ liệu: Một là thao tác dữ liệu trên khối lượng dữ liệu lớn, hai là khai phá độ đo phức tạp của dữ liệu.

Với đặc điểm này, cần thiết kế một mô hình gom cụm k nhóm theo trung bình của dữ liệu lớn hàng hải dựa trên kiến trúc Hadoop hiện thực mô hình MapReduce để xác định số lượng cụm tối ưu và tính toán độ lệch chuẩn của COG và SOG, cũng là vector đặc trưng khi tiến hành phân nhóm. Từ kết quả trực quan dữ liệu giúp cho người quản lý có thể nhận thức bằng thống kê và phân bố tàu biển tốt hơn. Các công việc thực hiện trong bài báo tối ưu dữ liệu hàng hải gom cụm $k$ nhóm trung bình dựa vào mô hình MapReduce bao gồm: Chọn trường dữ liệu hàng hải, tiền xử lý dữ liệu, thuật toán K-means, thống kê và trực quan hóa dữ liệu, kết luận và phản hồi. Được thể hiện trong Hình 2.



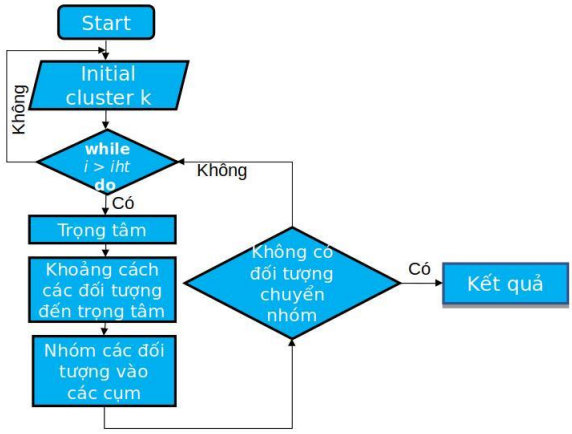

*Hình 2. Quy trình phân tích và trực quan hóa dữ liệu*

Theo quy trình tại bước tiền xử lý dữ liệu, là bao gồm phát hiện và loại bỏ lỗi dữ liệu, chuyển đổi định dạng và trích xuất dữ liệu nguồn. Sau bước tiền xử lý và chọn trường dữ liệu hàng hải, đến bước lựa chọn thuật toán K-means để thực hiện bước gom cụm tương ứng và thống kê và trực quan hóa dữ liệu. Thông qua việc trực quan hóa dữ liệu chúng tôi phân tích các kết quả để đưa ra kết luận, đồng thời lựa chọn nội dung nhằm nâng cao giá trị hiển thị thông tin hàng hải tàu biển.

## 2. Thuật toán K-means và kiến trúc Hadoop hiện thực mô hình MapReduce

### *2.1. Thuật toán K-means*

Thuật toán K-means [4] được sử dụng trong phân tích tính chất cụm của dữ liệu. Được thể hiện dưới Hình 3.




***Hình 3. Lưu đồ phân tích thuật toán K-means***

Phát biểu bài toán:

Input:

- Tập các đối tượng X = {$x_j$| j = 1, 2, ..., N}, $x_j \in R^d$

- Số cụm: k.

Output:

- Các cụm $C_i$ (i = 1 ÷ k) tách rời và hàm tiêu chuẩn E đạt giá trị tối thiểu.

- Thuật toán hoạt động trên 1 tập vector d chiều, tập dữ liệu X gồm N phần tử.

X = {$x_j$| j = 1, 2, ..., N}

- K-means lặp lại nhiều lần quá trình:

+ Gán dữ liệu;

+ Cập nhật lại vị trí trọng tâm.

- Quá trình dừng lặp lại khi trọng tâm hội tụ và mỗi đối tượng là một bộ phận của 1 cụm.

Hàm đo độ tương tự sử dụng khoảng cách Euclidean:

$$E = \sum_{j=1}^{N} \sum_{x_j \in c_i}^{k} \left( \|x_j - c_i\|^2 \right) \quad (1)$$

Trong đó, $c_i$ là trọng tâm của cụm $C_i$.

Thuật toán K-means chỉ định mỗi điểm $c_i$ cho cụm có trung tâm gần nhất, là giá trị trung bình cộng cho từng thứ nguyên riêng biệt trên tất cả các điểm trong cụm. Sau đây là các bước mô tả thuật toán K-means:

**Thuật toán 1: K-means (X, k)**

1. Với i = Float.MaxValue; j=1.
2. Chọn k là trung tâm tập X, với $C^{(0)} = c_1^{(j)}, c_2^{(j)}, ... c_k^{(j)}$.
3. while i > $i_{ht}$ do //với $i_{ht}$ là biên hội tụ.
4. k cụm (bằng cách gán mỗi điểm trung tâm gần nhất trong tập X).
5. Tìm điểm trung tâm mới của k cụm $c_1^{(++j)}, c_2^{(++j)}, ... c_k^{(++j)}$.
6. $i \leftarrow \sum_{m=0}^{k} \|c_m^j - c_m^{j-1}\|^2$ (2)
7. Kết quả $C^{(j)}$.

- Tính toán khoảng cách:

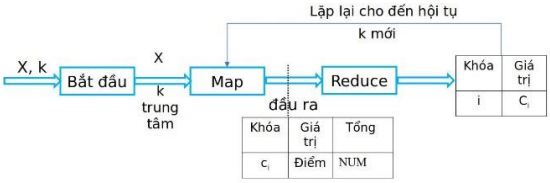
$$C_i^{(t)} = \{x_j: \|x_j - c_i^{(t)}\|^2 <= \|x_j - c_{i^*}^{(t)}\|^2 , i^* = 1, 2, ...k\} \quad (3)$$

- Cập nhập lại trọng tâm:

$$c_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{x_j \in C_i^{(t)}} x_j \quad (4)$$

### *2.2. Thuật toán MapReduce*

*a) Mô hình:* Được thể hiện trong Hình 4.




***Hình 4. Gom cụm K-means dựa vào mô hình MapReduce***

*b) Thuật toán:*

**- Thuật toán mapper:**

**Thuật toán 2: mapper (X, k)**

Ngõ vào: Biến trung tâm, khóa k, X là tập đối tượng.
Ngõ ra: <i, Điểm, NUM> , trong đó, i là điểm trung tâm gần nhất và Điểm là chuỗi giá trị thông tin.
1. Xây dựng kịch bản mẫu từ các giá trị X;
2. Khoảng cách dist = Double, giá trị lớn nhất;
3. Chỉ số index = -1;
4. For i = 0 to length.trung tâm do
Khoảng cách = hàm tính khoảng cách (mẫu kịch bản, trung tâm (i));
IF khoảng cách < khoảng cách dist {
khoảng cách dist = khoảng cách;
chỉ số index = i;
}
5. Kết thúc vòng lặp For;
6. Xây dựng giá trị Điểm từ chuỗi giá trị từ kịch bản;
7. Khởi tạo bộ đếm NUM để ghi các tổng số mẫu trong cùng một cụm;
8. While (Điểm.hasnext ()){
Xây dựng kịch bản từ Điểm.next();
Thêm giá trị kịch bản vào mảng;

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

NUM= NUM++;}
9. Ngõ ra: cặp <i, Điểm, NUM>;
10. Kết thúc mapper(X, k).

**- Thuật toán reducer:**

**Thuật toán 3: reducer (i, Điểm, NUM)**

Ngõ vào: Chỉ số cụm i, Tập hợp các giá trị Điểm, Tổng giá trị NUM;

Ngõ ra: <i, Ci> , trong đó, i là chỉ mục của cụm và Ci là giá trị trung tâm mới đại diện chuỗi.

1. Khởi tạo mảng các giá trị chứa cùng một cụm, ví dụ: Kịch bản trong danh sách Điểm;
2. Chia các mục của mảng cho NUM để có tọa độ điểm trung tâm;
3. Xây dựng giá trị một chuỗi bao gồm các tọa độ điểm trung tâm;
4. Ngõ ra cặp <i, Ci>;
5. Kết thúc reducer (i, Điểm, NUM).

Thuật toán sẽ dừng lại sau một số hữu hạn vòng lặp.

### *2.3. Kiến trúc Hadoop hiện thực mô hình MapReduce*

Chúng tôi sẽ tập trung vào kiến trúc Hadoop MapReduce, đây là cách triển khai mã nguồn mở phổ biến nhất hiện thực mô hình MapReduce do Google đề xuất. Kiến trúc Hadoop MapReduce chủ yếu bao gồm hai chức năng do người sử dụng xác định: map() và reduce(). Đầu vào của kiến trúc Hadoop MapReduce là cặp khóa - giá trị (k, v) và được gọi hàm map () cho mỗi cặp này. Hàm ánh xạ map () được tạo từ giá trị (0) hoặc nhiều cặp trung gian khóa - giá trị (k', v'). Sau đó, kiến trúc Hadoop MapReduce nhóm các cặp trung gian khóa-giá trị này bằng khóa trung gian k' và gọi là hàm reduce () cho mỗi nhóm. Cuối cùng, thuật toán reduce () được tạo ra giá trị (0) hoặc nhiều kết quả tổng hợp. Với kiến trúc Hadoop MapReduce chỉ sử dụng 2 tác vụ là định nghĩa hàm map () và reduce () để thực hiện phân tích dữ liệu quy mô lớn. Tuy nhiên, hiệu suất Input/ Output của kiến trúc Hadoop MapReduce phụ thuộc vào hệ thống phân tán Hadoop (HDFS), là bản sao mã nguồn mở của hệ thống Google.

Một trong những ưu điểm chính của kiến trúc Hadoop MapReduce là dùng máy tính thường chạy các tác vụ phân tích trên dữ liệu hàng hải lớn.

## 3. Gom cụm K-means dựa vào mô hình MapReduce

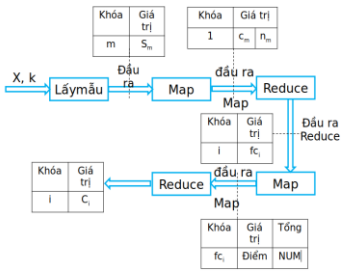
### *3.1. Phương pháp*

Gom cụm K-means dựa vào mô hình MapReduce được giả định cần thiết lập số lượng k cụm và lặp lại quá trình bằng cách di chuyển điểm trung tâm của cụm đến điểm trung bình của tập cụm dữ liệu cho đến khi điểm trung tâm của cụm hội tụ. Trong nghiên cứu này, chúng tôi chia nhỏ các phần tử bên trong nó, nghĩa là lấy mẫu từ tập dữ liệu đầu vào (X, k), với $n_m$ điểm thuộc tâm $c_m$, $fc_i$ đại diện cho tâm thứ $i$ cuối, với $i$ từ $1$ đến $k$; Quá trình này được mô tả trong Hình 5 dưới đây:

$$\text{argmin} \sum_{i=1}^{k} \sum_{x \epsilon C_i} (\|x - c_i\|^2) \quad (5)$$

$$c_i = \frac{1}{|C_i|} \sum_{x \epsilon C_i} x \quad (6)$$

Trong đó, $k$ là số cụm, $c_i$ là điểm trọng tâm *(trung tâm)* của cụm $C_i$, x là vector đặc trưng quỹ đạo của mỗi tàu biển.




***Hình 5. Tối ưu gom cụm K-means dựa vào mô hình MapReduce***

Để xác định số cụm $k$, chúng tôi dùng quy tắc khủy tay để tính số lượng gom cụm tối ưu. Tại bước đầu tiên, chúng tôi tính tổng khoảng các Euclid từ mọi mẫu đến điểm trung tâm của cụm và tiến hành các giá trị khác nhau của $k$. Tổng khoảng cách giảm khi $k$ tăng, vì vậy nó sẽ hội tụ. Bước tiếp theo, chúng tôi vẽ tại $k$ và tổng khoảng cách, vị trí của điểm lớn nhất (khủy tay) được xem là điểm hội tụ.

### *3.2. Thử nghiệm và phân tích*

Thông tin về giao thông hàng hải tàu biển trong khu vực miền Nam, Việt Nam rất phong phú. Quỹ đạo của các con tàu được xác định bằng cách liên kết thông tin vị trí của con tàu được hệ thống thông tin nhận dạng AIS thu thập gửi về trung tâm vận hành hệ thống. Tuy nhiên, lượng dữ liệu AIS thu thập của mỗi tàu không đồng đều, có thể tắc nghẽn tín hiệu hoặc hỏng máy phát tín hiệu nhận dạng và được xác định qua danh tính dịch vụ di động hàng hải (MMSI); Trong trường hợp này, chúng tôi loại bỏ thông tin của tàu thu thập, vì không đầy đủ thông tin hành trình của tàu. Độ lệch chuẩn x = (speed, course) (*(tốc độ, hướng)*), với 'course' được đổi từ độ (°) sang radian

trước khi chuẩn hóa, của mỗi tàu bằng vector đặc trưng $x_{chuẩn\ hóa}$ = (SOG_lc, COG_lc) để đánh giá mức độ ổn định của tàu. Và chuẩn hóa mẫu trước khi gom cụm, bằng logarit như phương trình:

$$x_{chuẩn\ hóa} = \log_{10}(x+1)/\log_{10}(x_{max}+1) \quad (7)$$

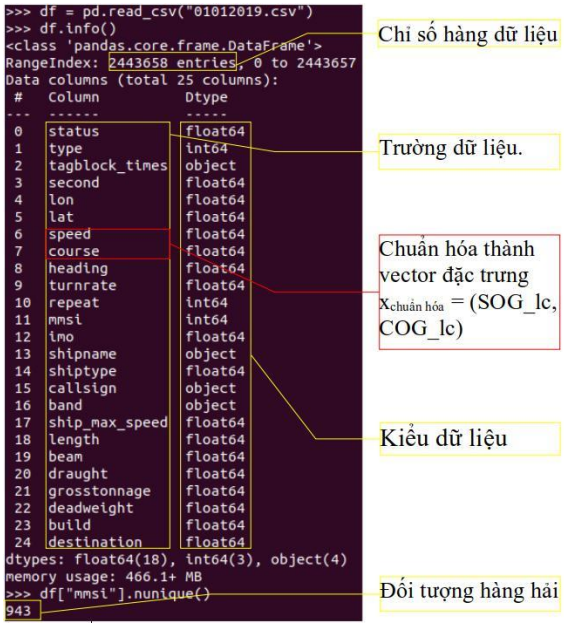
*3.2.1. Dữ liệu*

Tập dữ liệu thu thập AIS:

| Thời gian | Tập dữ liệu *(số mẫu dữ liệu)* | Đối tượng hàng hải | Trường dữ liệu |
|---|---|---|---|
| 13/9/2019 | 157.773.900 | 1089 | 25 |

* Đối tượng hàng hải *(tàu, thiết bị báo hiệu tích hợp AIS,..);*

** Trường dữ liệu *(gồm thông tin di động, tĩnh của đối tượng hàng hải)*.



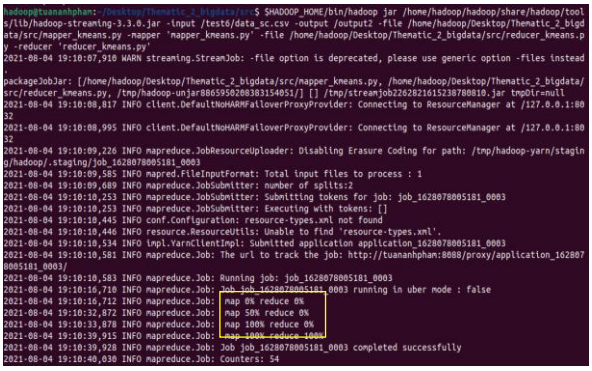

***Hình 6. Thông tin chi tiết khung dữ liệu AIS thu thập***

*3.2.2. Thực hiện chạy gom cụm K-mean dựa vào mô hình MapReduce*

a) Lấy ngẫu nhiên 3 điểm trung tâm và kết quả khi chạy mô hình:

| k | SOG_lc (speed) | COG_lc (course) |
|---|---|---|
| 1 | 0.400000005960465 | 4.09999990463257 |
| 2 | 8.69999980926514 | 4.0 |
| 3 | 0.0 | 141.899993896484 |
| 4 | 0.0 | 233.1999969482424 |
| 5 | 0.0 | 187.0 |

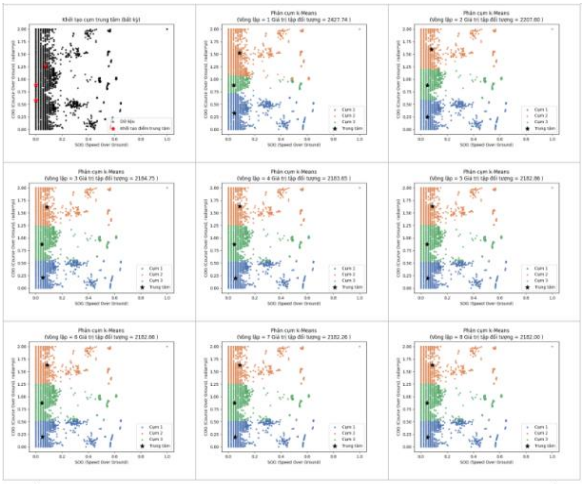
- Chạy mô hình:

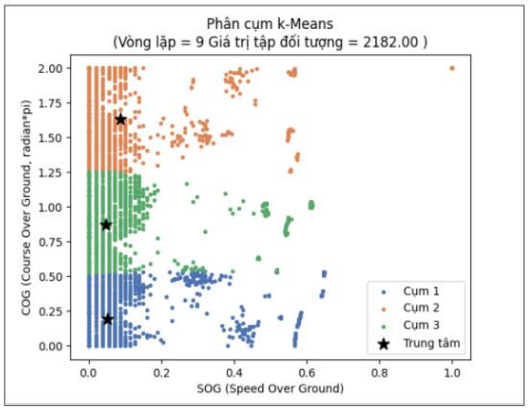



***Hình 7. Thực thi gom cụm K-means dựa vào mô hình MapReduce ứng với k = 5***

- Kết quả gom cụm:

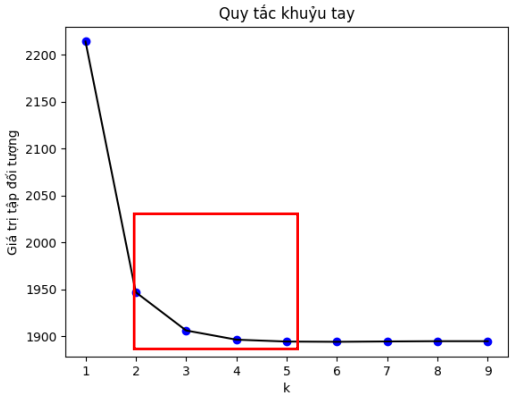
| k | SOG_mới (speed) | COG_mới (course) |
|---|---|---|
| 1 | 0.1515723280061966 | 40.15723284380803 |
| 2 | 10.542187556624407 | 5.099999967962503 |
| 3 | 1.8082741391924422 | 126.4260651983998 |
| 4 | 5.346773882370912 | 286.8167981761306 |
| 5 | 0.8841017462988348 | 193.6729736813302 |






***Hình 8. Kết quả gom cụm K-means được chuẩn hóa ứng k = 5***

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

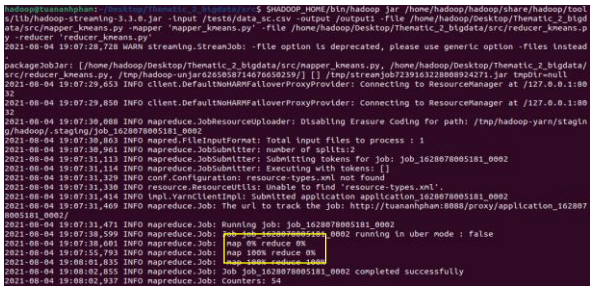



*Hình 9. Giá trị hàm mục tiêu ứng với k = 5*

b) Lấy ngẫu nhiên 8 điểm trung tâm và kết quả khi chạy mô hình:

| k | SOG_lc (speed) | COG_lc (course) |
|---|---|---|
| 1 | 0.400000005960465 | 4.09999990463257 |
| 2 | 8.69999980926514 | 4.0 |
| 3 | 0.0 | 141.899993896484 |
| 4 | 0.0 | 233.199996948242 |
| 5 | 0.0 | 187.0 |
| 6 | 12.0 | 187.100006103516 |
| 7 | 7.80000019073486 | 341.0 |
| 8 | 0.0 | 304.200012207031 |

- Chạy mô hình:

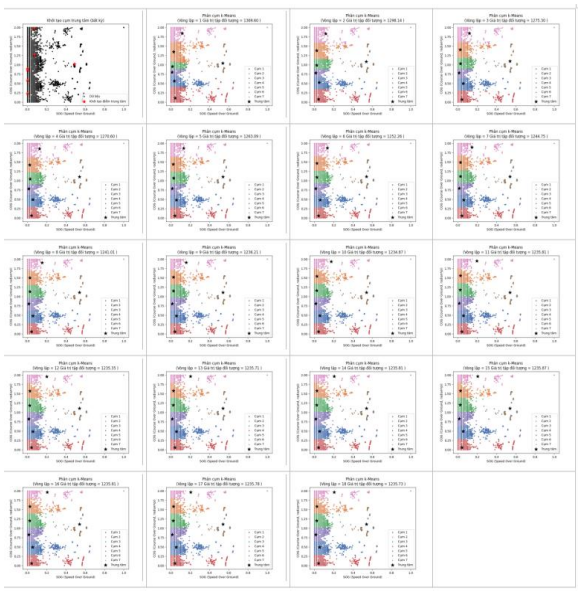


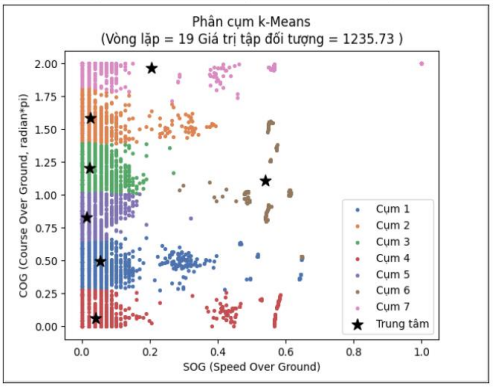
*Hình 10. Thực thi gom cụm K-means dựa vào mô hình MapReduce ứng với k = 8*

- Kết quả gom cụm:

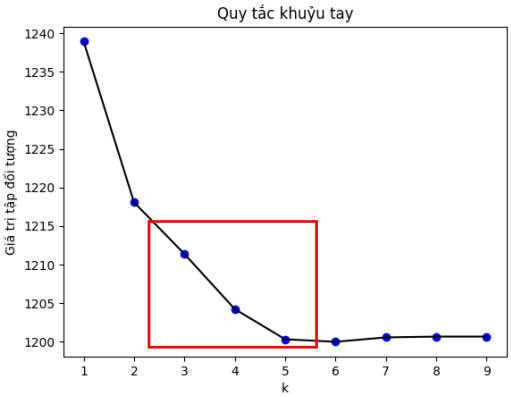
| k | SOG_mới (speed) | COG_mới (course) |
|---|---|---|
| 1 | 0.1515723280061966 | 40.15723284380803 |
| 2 | 10.542187556624407 | 5.099999967962503 |
| 3 | 1.8082741391924422 | 126.4260651983998 |
| 4 | 0.2811068720433092 | 234.49637463984598 |
| 5 | 0.10534482568759343 | 193.94638006276097 |
| 6 | 10.10204080659516 | 190.43673488071985 |

| k | SOG_mới (speed) | COG_mới (course) |
|---|---|---|
| 7 | 14.876438070975322 | 349.73714159395746 |
| 8 | 0.6731448789578027 | 290.9056507555419 |






*Hình 11. Kết quả gom cụm K-means được chuẩn hóa ứng k = 8*




*Hình 12. Giá trị hàm mục tiêu ứng với k = 8*

Dựa vào kết quả thu được, cùng với hàm giá trị mục tiêu (ứng với k = 5, k = 8) chúng tôi xác định được số lượng cụm tối ưu (k = 5). Qua đó, chúng tôi thu hẹp phạm vi tính chất cụm để đánh giá hàng hải tàu biển.

## 4. Kết luận

Trong bài nghiên cứu này, chúng tôi đã thực hiện phân tích dữ liệu hàng hải lớn bằng phương pháp gom cụm K-means dựa vào mô hình MapReduce để xử lý các đặc trưng dữ liệu *(độ lệch chuẩn của cặp (speed, course))* của mỗi tàu biển từ hệ thống nhận dạng tự động AIS được thu thập gửi về trung tâm vận hành hệ thống. Với phương pháp này, người vận hành hệ thống có thể giám sát, đánh giá được sự ổn định giao thông hàng hải tàu biển và dùng để phát hiện sự bất thường trong hàng hải tàu biển dựa vào tính chất của cụm. Tuy nhiên, thông tin AIS thu thập còn có nhiều đặc trưng mà chúng tôi chưa khai thác hết, các đặc trưng trích xuất vẫn còn đơn giản và có thể nâng cao cho các ấn phẩm tiếp theo trong tương lai.

## TÀI LIỆU THAM KHẢO

[1] Hadoop: *Open source implementation of MapReduce*, https://hadoop.apache.org/.

[2] Phạm Tuấn Anh, *Đưa công nghệ vào bảo đảm an toàn hàng hải luồng Vũng Tàu - Thị Vải*. Tạp chí Giao thông vận tải, 2019, http://www.tapchigiaothong.vn/.

[3] Automatic identification systems (AIS). IMO, https://www.imo.org.

[4] Jiawei Han, Micheline Kamber, Jian Pei, *DataMining: Concepts and Techniques*, Third Edition, Morgan Kaufmann Publishers, 2011.

Ngày nhận bài: 05/10/2021
Ngày nhận bản sửa: 15/10/2021
Ngày duyệt đăng: 23/10/2021