

GOM CỤM QUỸ ĐẠO ĐƯỜNG ĐI DỰA VÀO TỐC ĐỘ VÀ CẢI TIẾN BẰNG PHƯƠNG PHÁP RÚT TRÍCH ĐẶC TRƯNG DỰA TRÊN HISTOGRAM

CLUSTERING TRAJECTORIES BASED ON SPEED FEATURES AND AN IMPROVEMENT BY USING HISTOGRAM – BASED FEATURE EXTRACTION

Nguyễn Công Hà, Lê Văn Quốc Anh

Khoa CNTT, ĐH GTVT TP.HCM

conghanguyen@gmail.com

Tóm tắt: Gom cụm quỹ đạo từ dữ liệu hành trình GPS để tìm các nhóm hành trình tương tự nhau, từ lâu đã được xem là một nhiệm vụ quan trọng để phân đoán các luồng giao thông chính, từ đó cho phép dự đoán vị trí hay hành vi di chuyển của các phương tiện giao thông. Đa số các phương pháp hiện có đều dùng đặc trưng gom cụm dựa trên không gian và thời gian nhưng những đặc trưng này là không phù hợp để phân loại hành vi di chuyển của các phương tiện. Trong bài báo này nhóm nghiên cứu tập trung vào vấn đề gom cụm hành trình phương tiện có chung hành vi di chuyển. Ví dụ hành vi di chuyển của xe ô tô cá nhân thì khác xe buýt và cũng khác với xe taxi. Giải pháp mà nhóm đưa ra là chuyển đổi từng quỹ đạo thành một chuỗi đặc trưng để mô tả các hành vi chuyển động của đối tượng và áp dụng các giải thuật gom cụm trên không gian đặc trưng. Nhóm thực nghiệm trên các bộ dữ liệu thực và thu được kết quả tốt hơn các giải pháp hiện có.

Từ khóa: Dữ liệu hành trình GPS; khai thác dữ liệu; phát hiện hành vi di chuyển.

Chỉ số phân loại: 1.4

Abstract: Clustering trajectories from GPS data to find groups of similar trajectories is a crucial task for recognizing main traffic flows and then to predict next positions as well as moving behavior of vehicles. Most existing approaches are based on features extracted from raw locations and times. Such features are not suitable for classifying moving behaviors of vehicles. In this paper, we focus on clustering trajectories of vehicles having the same moving behaviors. For example, moving behaviors of cars is different from moving behaviors of buses or taxi. Our proposed approach is transforming each trajectory to a sequence of features to model moving behaviors of an object and utilizing traditional clustering algorithms to groups trajectories. We experiment on real datasets and obtain better results than existing approaches.

Keywords: GPS trajectory data; data mining; behavior detection.

Classification number: 1.4

1. Giới thiệu

Ngày nay, sự phổ biến rộng khắp và nhanh chóng của các thiết bị được trang bị GPS hay các dịch vụ dựa trên vị trí đã tạo ra một khối lượng dữ liệu hành trình lớn của các đối tượng chuyển động. Trong phạm vi các phương tiện giao thông, kho dữ liệu hành trình như vậy là rất đáng giá cho việc khai thác các thông tin hữu ích, như phát hiện các điểm ùn tắc hay các luồng di chuyển chính trong mạng lưới giao thông. Trong bài báo này, nhóm nghiên cứu tập trung vào hướng phân tích dữ liệu hành trình của phương tiện dựa trên các giải thuật gom cụm. Việc khám phá các cụm quỹ đạo không chỉ tiết lộ các đặc điểm hành vi của các phương tiện chuyển động mà cao hơn nữa là các dự đoán về hành trình, dự đoán bất

thường hay đưa ra các mẫu hành vi di chuyển. Hiện nay có nhiều kỹ thuật gom cụm hành trình GPS đã được đề xuất, chẳng hạn [1], [2], [3]. Những giải pháp này sử dụng các biện pháp nhất định để định nghĩa tính tương đồng về quỹ đạo và sau đó áp dụng vào các thuật toán phân cụm cổ điển. Mặc dù các biện pháp này có thể nhóm các quỹ đạo tương tự trong một khu vực địa lý cố định và khoảng thời gian, nhưng để gom các quỹ đạo dựa trên hành vi di chuyển tương tự nhau thì cần phải có một giải pháp mới. Với các phương tiện giao thông, hành vi di chuyển của từng loại phương tiện là khác nhau. Ví dụ xe ô tô cá nhân thì thường ghé những địa điểm cố định (nhà, cơ quan, trung tâm mua sắm,...), xe buýt thường đi theo những lộ trình và ghé ở các trạm cố

định, trong lúc xe taxi thì số địa điểm ghé là khá phong phú. Trong bài báo này, nhóm đề xuất giải pháp trích xuất đặc trưng từ hành trình GPS để đưa vào giải thuật gom cụm hành vi di chuyển. Ngoài giải pháp thông thường dựa trên đặc trưng là các giá trị thống kê từ chuỗi tốc độ, nhóm đề xuất dùng kỹ thuật dựa trên histogram để trích xuất đặc trưng. Nhóm thử nghiệm trên hai bộ dữ liệu thật. Một bộ là dữ liệu lấy từ kho dữ liệu có tiếng UCI Machine Learning, với các hành trình có gán nhãn. Nhóm dùng dữ liệu này để đánh giá độ chính xác của giải thuật và nhận thấy giải pháp được nhóm đề xuất cho kết quả tốt hơn nhiều so với dùng các giá trị thống kê làm đặc trưng. Ngoài ra, nhóm dùng bộ dữ liệu thứ hai gắn với đường phố của Việt Nam, để chứng tỏ tính hữu dụng khi áp dụng vào các thành phố ở Việt Nam.

2. Các khái niệm và công trình liên quan

2.1. Mô hình hoá dữ liệu GPS

Nhóm sử dụng các khái niệm sau đây để mô hình hoá dữ liệu hành trình GPS:

- **Toạ độ GPS:** Được biểu diễn bởi bộ bốn $\langle id, lat, lon, time \rangle$, trong đó: id là mã định danh của đối tượng chuyển động (phương tiện giao thông hoặc một điện thoại có hỗ trợ GPS); lat là vĩ độ, lon là kinh độ; và time là thời gian ghi nhận vị trí của đối tượng.
- **GPS Log:** Là một tập hợp các toạ độ GPS, có dạng $\{p_1, p_2, \dots, p_n\}$, với mỗi p_i là một toạ độ GPS.
- **Hành trình GPS:** Là một chuỗi gồm các toạ độ GPS thu thập không ngắt quãng của cùng một đối tượng chuyển động, có dạng $p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n$.

Dữ liệu đầu vào cho các giải thuật gom cụm sẽ là các hành trình GPS được trích xuất từ GPS Log, đó là dữ liệu thô. Kết quả, như phần thực nghiệm sẽ trình bày.

2.2. Tình hình nghiên cứu gần đây

Các phương pháp gom cụm hành trình cổ điển áp dụng các thuật toán gom cụm dựa trên khoảng cách hoặc dựa trên mật độ. Do các hành trình GPS thường có độ dài biến động nên thông thường, hành trình được tách thành các hành trình con và sau đó gom nhóm hành

trình con bằng cách sử dụng các phương pháp gom cụm dựa trên mật độ [1]. Ngoài ra, một số kỹ thuật khác đề xuất như kết hợp lấy mẫu với gom cụm dựa trên mật độ để xử lý nhiễu trong dữ liệu quỹ đạo [2]. Đối với nhóm các giải pháp gom cụm dựa trên khoảng cách, các giải pháp này cố gắng định nghĩa độ đo khoảng cách hoặc độ đo tương tự giữa hai hành trình dựa trên các hàm tính khoảng cách hiện có như Eclidean hay DTW [3]. Một số kỹ thuật định giới thiệu một phép đo khoảng cách mới [4]. Khác với các phương pháp trên, nhóm không gom cụm trên dữ liệu gốc mà biến đổi sang một không gian đặc trưng. Các giải thuật gom cụm sẽ thực thi trên không gian đặc trưng mới này.

3. Nguồn dữ liệu thực nghiệm

Nguồn dữ liệu thực nghiệm gồm có hai bộ dữ liệu. Bộ dữ liệu thứ nhất lấy từ kho dữ liệu UCI Machine Learning

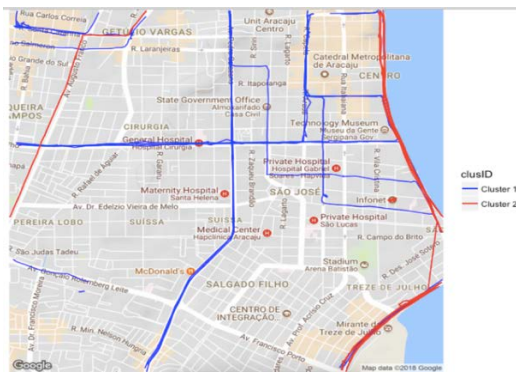
(<https://archive.ics.uci.edu/ml/datasets/GPS+Trajectories>), thu thập từ một ứng dụng Android gọi là Go!Track. Dữ liệu này có 163 hành trình, thời gian thu thập dữ liệu 29/02/2016. Mỗi hành trình đều có gán sẵn nhãn là car_or_bus để phân biệt hành trình này là của xe ô tô hay xe buýt. Bộ dữ liệu thứ hai là các hành trình di chuyển của phương tiện ở Việt Nam được cung cấp bởi công ty OTS cung cấp giám sát dịch vụ hành trình xe ô tô. Số lượng xe được giám sát hành trình trong nguồn dữ liệu này là 411, khu vực Thành phố Hồ Chí Minh. Thời gian thu thập dữ liệu trong vòng một tuần từ 01/06/2015 đến 07/06/2015.

4. Khai thác thông tin hành vi di chuyển từ dữ liệu GPS

4.1. Trích xuất hành vi di chuyển dựa trên giá trị thống kê

Trước tiên nhóm sử dụng phương pháp trích xuất hành vi đối tượng bằng cách sử dụng các giá trị thống kê dựa trên tốc độ, tương tự giải pháp đề nghị bởi W. Tang và cộng sự [2]. Theo đó, chúng tôi lấy sáu định lượng {trung bình, lớn nhất, 75% định lượng, 50% định lượng, 25% định lượng, nhỏ nhất} để làm đặc trưng cho các hành trình. Một giải thuật gom cụm bất kỳ được áp dụng trên không gian đặc trưng để gom cụm các hành trình. Để kiểm nghiệm độ chính xác, nhóm dùng bộ dữ liệu

đã có gán nhãn và cố gắng phân thành các cụm, số lượng cụm tương ứng với số lượng nhãn khác nhau. Vì biết trước số cụm nên nhóm chọn giải thuật K-MEANS để thử nghiệm. Với bộ dữ liệu cho biết trước gồm các hành trình xe ô tô và xe buýt đã giới thiệu ở phần trên thì chúng tôi gom thành hai cụm và đánh giá độ chính xác. Kết quả là độ chính xác 64% cho hai cụm là không cao. Xem xét các hành trình trên bảng đồ thì thấy các hành trình của xe ô tô và xe buýt gán nhầm vào cùng một cụm do chúng cùng chạy trên một tuyến đường, như minh họa hình 1.



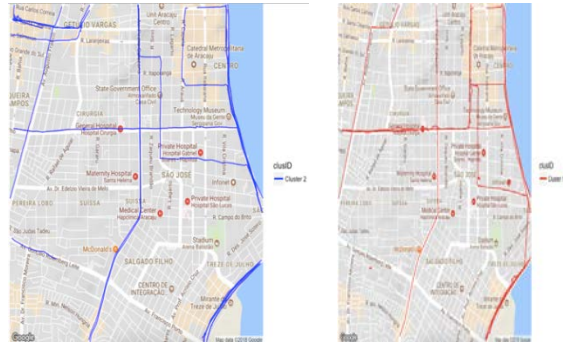
Hình 1. Kết quả K-Means tìm hành vi di chuyển.

Rõ ràng là cách thực hiện như vừa rồi không thích hợp cho việc phân loại hành trình dựa trên hành vi di chuyển. Để cải tiến giải pháp này, nhóm đề xuất phương pháp trích xuất đặc trưng dùng Histogram.

4.2. Cải tiến bằng phương pháp trích xuất đặc trưng dựa vào Histogram

Qua quá trình thực nghiệm nhóm nhận ra rằng có thể gom cụm tốt có độ chính xác cao hơn với đặc trưng dựa vào histogram thay vì các giá trị thống kê trên mật độ. Nhóm mô phỏng hành vi di chuyển với các ngưỡng 0, 5, 10, 15, 20, 25, 30, 40, 60, 100 (km/h) và dùng các ngưỡng này xác định các bin để tính histogram. Histogram sau khi tính được sẽ thành đặc trưng trong không gian chín chiều. Tương tự như phần trên, nhóm áp dụng giải thuật gom cụm K-MEANS trên không gian này với số cụm thiết lập là 2. Độ chính xác của phương pháp này là 96%.

Hình 2 minh họa kết quả của giải pháp này, với các xe ô tô và xe buýt chạy cùng tuyến đường vẫn được phân cụm đúng.

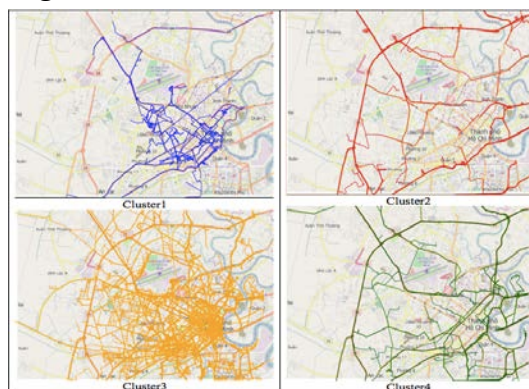


Hình 2. Kết quả gom cụm hành trình: Cluster 1 gồm các hành trình của xe ô tô và Cluster 2 chứa các hành trình xe buýt.

4.3. Thực nghiệm phương pháp histogram vào dữ liệu Việt Nam

Phần này nhóm áp dụng giải pháp đề xuất trên dữ liệu GPS của Việt Nam. Đây là bộ dữ liệu thứ hai đã giới thiệu ở mục 3. Ngoài giải pháp gom cụm K-MEANS, nhóm còn thử trên giải thuật gom cụm dựa trên mật độ DBSCAN. Kết quả thực nghiệm của hai giải thuật không khác nhau nhiều, do bản chất ở đây là kỹ thuật trích xuất đặc trưng trước khi đưa vào giải thuật gom cụm.

Hình 3 minh họa các hành trình sau khi được gom thành bốn cụm. Với cụm 1 là các xe cá nhân di chuyển với các tuyến đường chính bên trong thành phố và thỉnh thoảng sử dụng các đường quốc lộ. Cụm số 2 là các xe khách đường dài, tuyến đường di chuyển chính là xa lộ và các trục đường chính trong nội đô. Cụm số 3 là các xe buýt, di chuyển ở các tuyến cố định. Đặc biệt cụm số 4 di chuyển hầu hết các tuyến đường, đây là hành trình của các xe taxi. Với phần thực nghiệm trên, phương pháp này cho thấy có thể ứng dụng tốt trong việc phân loại hành vi di chuyển của phương tiện giao thông.



Hình 3. Thực nghiệm phương pháp Histogram vào dữ liệu Việt Nam.

5. Kết luận và hướng phát triển

Trong bài báo này nhóm đề xuất một phương pháp mới cho gom cụm dữ liệu hành trình GPS có sự tương tự nhau về hành vi di chuyển. Hành vi di chuyển được trích xuất từ không gian và thời gian bất biến của quỹ đạo với việc sử dụng tốc độ để theo dõi hành vi của đối tượng. Đề xuất mới là cải tiến bằng phương pháp Histogram đưa lại kết quả chính xác và hiệu quả cao hơn so với gom cụm dựa vào các giá trị thống kê trên tốc độ □

Lời cảm ơn

Nghiên cứu này được hỗ trợ từ nguồn kinh phí nghiên cứu khoa học của Trường Đại học Giao thông vận tải Thành phố Hồ Chí Minh (MS KH1705).

Tài liệu tham khảo

- [1] J. H. Jae-Gil Lee and Kyu-Young Whang, "Trajectory Clustering: A Partition-and-Group Framework," in *SIGMOD'07*, 2007, pp. 593–604.

[2] W. Tang, D. Pi, and Y. He, "A density-based clustering algorithm with sampling for travel behavior analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9937 LNCS, pp. 231–239, 2016.

- [3] Z. Li, J. G. Lee, X. Li, and J. Han, "Incremental clustering for trajectories," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 5982 LNCS, no. PART 2, pp. 32–46.

- [4] G. Yuan, P. Sun, J. Zhao, D. Li, and C. Wang, "A review of moving object trajectory clustering algorithms," *Artif. Intell. Rev.*, vol. 47, no. 1, pp. 123–144, 2017.

Ngày nhận bài: 29/6/2018

Ngày chuyển phản biện: 3/7/2018

Ngày hoàn thành sửa bài: 24/7/2018

Ngày chấp nhận đăng: 31/7/2018