

Sử dụng thuật toán K-means trong bài toán phân loại đám mây điểm LiDAR

Nguyễn Thị Hữu Phương^{1*}, Nguyễn Trường Xuân¹, Đặng Văn Đức²

¹Trường Đại học Mỏ - Địa chất

²Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Ngày nhận bài 7/3/2017; ngày chuyển phản biện 10/3/2017; ngày nhận phản biện 7/4/2017; ngày chấp nhận đăng 17/4/2017

Tóm tắt:

Hiện nay, LiDAR (Light detecting and ranging) là một công nghệ viễn thám mới đang được ứng dụng khá rộng rãi trong nhiều ngành, nhiều lĩnh vực. Xử lý dữ liệu LiDAR là bài toán không dễ dàng, trong khi các phần mềm xử lý dữ liệu LiDAR thường là phần mềm đóng và có chi phí khá cao. Bài báo đề cập đến việc phân loại dữ liệu LiDAR sử dụng thuật toán K-means, một bước tương đối quan trọng trong xử lý dữ liệu LiDAR, nhằm giúp phân chia các điểm về các lớp của nó, từ đó có thể ứng dụng vào các bài toán khác nhau trong thực tế.

Từ khóa: Đám mây điểm LiDAR, K-means, LiDAR, phân loại.

Chỉ số phân loại: 1.2

Using K-means algorithms for LiDAR point cloud classification

Thi Huu Phuong Nguyen^{1*}, Truong Xuan Nguyen¹,
Van Duc Dang²

¹Hanoi University of Mining and Geology

²Institute of Information Technology, Vietnam Academy of Science and Technology

Received 7 March 2017; accepted 17 April 2017

Abstract:

Nowadays, LiDAR is a technology applied widely for many fields and sectors. LiDAR data processing is not an easy issue, while LiDAR data processing software is usually closed and quite expensive. This paper deals with the classification of LiDAR data using K-means, an important step in LiDAR data processing, to divide the points to its classes so that they can be applied to different problems.

Keywords: Classification, K-means, LiDAR, LiDAR point cloud.

Classification number: 1.2

Đặt vấn đề

LiDAR là công nghệ viễn thám mới, chủ động, sử dụng các loại tia laser để khảo sát đối tượng từ xa. Dữ liệu thu được của hệ thống là tập hợp đám mây điểm phân xạ 3 chiều của tia laser từ đối tượng được khảo sát. Hiện nay, công nghệ LiDAR đang được ứng dụng rộng

rãi trong việc khảo sát địa hình và lập bản đồ, đánh giá sản lượng gỗ trong lâm nghiệp, lập bản đồ ngập úng, địa hình đáy biển, các tuyến truyền tải, bản đồ giao thông, mạng điện thoại di động, mô phỏng mô hình đô thị... và có tiềm năng trong nhiều ứng dụng khác như mô phỏng tác động của bão, tạo mô hình 3 chiều thành phố ảo, mô phỏng thiệt hại của động đất, khai khoáng, môi trường...

Hệ thống LiDAR là một hệ thống tích hợp từ 3 thành phần chính: Hệ thống thiết bị laser, hệ thống định vị vệ tinh GNSS và hệ thống đạo hàng quán tính INS. Ở mỗi thời điểm phát xung laser, hệ thống định vị vệ tinh GNSS sẽ xác định vị trí không gian của điểm phát, và hệ thống đạo hàng quán tính sẽ xác định các góc định hướng trong không gian của tia quét. Một tín hiệu phát đi sẽ có một hay nhiều tín hiệu phản xạ. Kết quả cuối cùng sẽ có được đám mây điểm. Để sử dụng các đám mây điểm cho mục đích thành lập mô hình số độ cao (Digital elevation model - DEM), mô hình số địa hình (Digital terrain model - DTM) hay mô hình số bề mặt (Digital surface models - DSM), phải tiến hành phân loại điểm trong đám mây điểm đó. Hiện nay, có nhiều thuật toán lọc điểm được sử dụng; với từng thuật toán, các hãng cung cấp thiết bị đã xây dựng phần mềm kèm theo trong một chu trình sử dụng đã được bảo mật. Để có thể phát huy hiệu quả của công nghệ LiDAR trong công tác trắc địa - bản đồ, việc hiểu biết sâu sắc về công nghệ và phát triển được các thuật toán phân loại điểm dữ liệu LiDAR đóng vai trò quan trọng [1].

Trên thế giới, việc phân loại dữ liệu LiDAR để từ đó trích xuất ra được các đối tượng phục vụ trong công tác

*Tác giả liên hệ: nguyenthphuong85.nb@gmail.com

xây dựng bản đồ và nhiều lĩnh vực khác của đời sống xã hội đã khá phổ biến. Trong các nghiên cứu [2-4] đã sử dụng các thuật toán phân loại để tiến hành phân loại đám mây điểm LiDAR, từ đó thành lập DTM, DSM, DEM và đã có những thành công nhất định.

Tại Việt Nam, việc phân loại dữ liệu LiDAR chủ yếu được tiến hành thủ công, hầu như chưa có công trình nghiên cứu cụ thể nào đề cập đến bài toán phân loại đám mây điểm LiDAR. Nghiên cứu của Trần Đình Luật [5] đã có một số kết quả thực nghiệm ban đầu, với địa hình tại các khu vực đảo Hòn Dấu, khu vực Vũng Tàu, Cần Giờ và các khu vực cửa sông ở Đồng bằng sông Cửu Long, kết quả quét LiDAR và thành lập DEM là khả quan. Nghiên cứu của tác giả Lương Chính Kế [6] và Trần Đức Phú [7] đã đề cập đến việc sử dụng dữ liệu LiDAR để phục vụ cho nhiều lĩnh vực khác nhau. Tuy nhiên, những nghiên cứu này chỉ sử dụng dữ liệu LiDAR sau khi đã được phân loại, sử dụng mô hình DEM có sẵn. Chúng tôi đã tiến hành nghiên cứu sử dụng thuật toán K-means trong bài toán phân loại đám mây điểm LiDAR nhằm tìm ra phương pháp phát huy hiệu quả của công nghệ LiDAR trong công tác trắc địa - bản đồ.

Nội dung nghiên cứu

Bài toán phân loại dữ liệu

Phân loại dữ liệu là quá trình tổ chức dữ liệu theo thể loại có liên quan để có thể sử dụng và bảo vệ dữ liệu hiệu quả hơn. Phân loại dữ liệu đặc biệt quan trọng khi nói đến quản lý rủi ro, tuân thủ và bảo mật dữ liệu [8]. Hướng tiếp cận này thường sử dụng một số kỹ thuật của học máy như cây quyết định, mạng nơ ron nhân tạo, học sâu...

Quá trình phân loại dựa trên 5 thành phần cơ bản:

- Bản ghi (Record).
- Lớp (Class).
- Dự đoán (Predictors).
- Tập dữ liệu huấn luyện (Training dataset).
- Tập dữ liệu kiểm tra (Testing dataset).

Đặc trưng của tiến trình phân loại gồm những điểm sau:

- Đầu vào: Tập dữ liệu đào tạo chứa những đối tượng với thuộc tính của nó, với một số thuộc tính đã được gán nhãn.

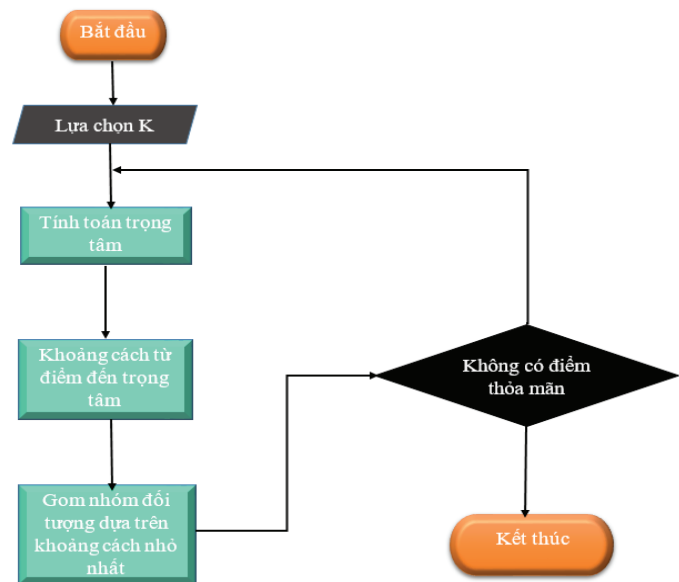
- Đầu ra: Mô hình phân lớp (Classifier) được gán bởi những nhãn cụ thể cho mỗi đối tượng (phân lớp các đối tượng theo từng thư mục), dựa trên những thuộc tính của

mỗi đối tượng.

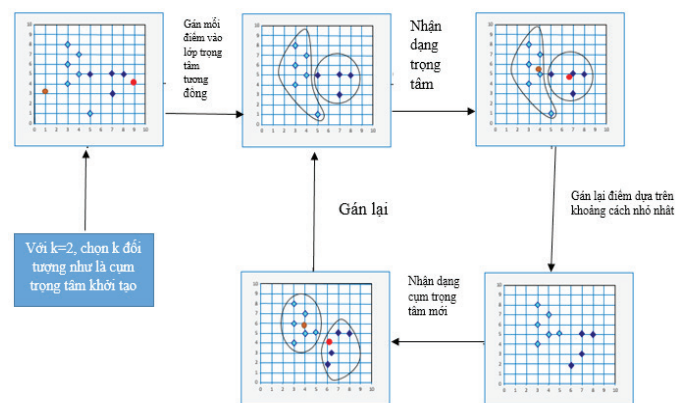
- Mô hình sử dụng để dự đoán những lớp mới, những đối tượng chưa biết. Tập dữ liệu kiểm thử cũng dùng để xác định độ chính xác của mô hình.

K-means trong bài toán phân loại

Thuật toán K-means là tìm phương pháp phân nhóm các đối tượng (Objects) đã cho vào K cụm (K là số cụm được xác định trước, $K > 0$) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm là nhỏ nhất. Thuật toán K-means được mô tả trên hình 1 và 2.



Hình 1. Mô tả thuật toán K-means.



Hình 2. Ví dụ thuật toán K-means.

Thuật toán K-means trong bài toán phân loại dữ liệu

Trong bài toán phân loại dữ liệu, thuật toán K-means được triển khai theo các bước sau [9-11]:

Bước 1: Chọn K cụm trọng tâm khởi tạo, $z_1, z_2, z_3, \dots, z_n$.

Bước 2: Phân phối mẫu trong K-means. Mẫu thường được gán với cụm trung tâm gần nhất theo công thức: $x \in S_i(n)$ nếu $|x - z_i(n)| \leq |x - z_j(n)|$ với $j = 1, 2, 3, \dots, k$; $i \neq j$; $S_i(n)$ là bộ mẫu của trọng tâm $z_i(n)$, trong đó n là chỉ số bước lặp của bài toán.

Bước 3: Tính toán trọng tâm cụm mới từ mỗi cụm $S_i(n)$. Tìm giá trị mới cho mỗi z_i . Trọng tâm cụm mới, $z_i(n+1)$ sẽ là giá trị trung bình của các điểm trong $S_i(n)$ như:

$$z_i(n+1) = (1/c_i) \sum_{x \in S_i(n)} x$$

Trong đó, c_i là tập điểm thuộc về cụm thứ i .

Bước 4: So sánh $z_i(n)$ và $z_i(n+1)$ với mọi i .

Tính toán khoảng cách giữa mỗi cặp điểm trong mỗi lần lặp liên tiếp:

a. Nếu không có sự thay đổi đáng kể, kết thúc phương pháp, một vài tiêu chí cho kết thúc như:

+ Nếu $|z_i(n+1) - z_i(n)| < T$ với mọi i

+ Nếu $\sum_{j=1}^k |z_j(n+1) - z_j(n)| < T$ với mọi i

b. Nếu không thì tiếp tục lặp các lần lặp tiếp theo từ bước 2.

K-means trong phân loại đám mây điểm LiDAR

Đặc điểm đám mây điểm LiDAR: Kết quả thu được sau khi xử lý dữ liệu không gian LiDAR gọi là đám mây điểm. Đám mây điểm đầu tiên là tập hợp các điểm độ cao, với tọa độ x, y cùng với bổ sung các thuộc tính như thời gian GPS. Các đặc trưng bề mặt được tia laser thu nhận được sau quá trình xử lý, ví dụ độ cao mặt đất, tòa nhà, tán cây, cầu vượt, và các đối tượng khác trong suốt quá trình quét được tín hiệu laser thu nhận được tạo thành đám mây điểm [1].

Một số thông số đặc trưng của đám mây điểm LIDAR:

- Tọa độ X, Y và độ cao Z : Được thu nhận dựa theo hệ thống định vị GPS, độ cao máy bay, thời gian di chuyển và phản xạ trở lại của tia laser...

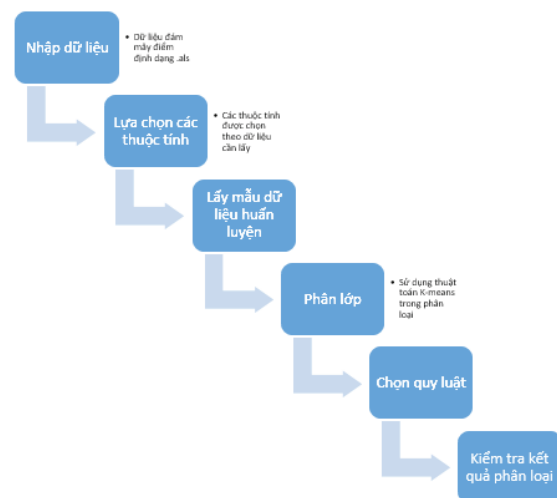
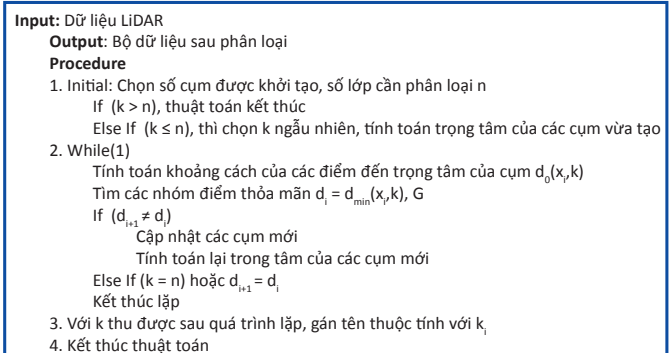
- Số lần phản xạ (Return): Các chùm tia laser sau khi chạm vào các đối tượng như tòa nhà, mặt đất, cột điện thì phản xạ (Return) ngược trở lại và được bộ thu nhận tín hiệu laser thu lại.

- Cường độ xung phản xạ (Intensity): Khi tia laser phản xạ trở lại sẽ mang theo năng lượng với một cường độ nhất định. Thông thường, cường độ xung phản xạ lớn khi tia laser tiếp xúc với mặt đất.

Ngoài ra còn có các thuộc tính của đám mây điểm

LIDAR như: Số phản hồi, góc máy bay, thời gian GPS góc quét, hướng quét...

Phân loại đám mây điểm LiDAR sử dụng K-means: Mỗi điểm LiDAR trong quá trình phân loại được gán vào một lớp được định nghĩa trong quá trình phân loại. Các điểm này có thể được phân vào một số lớp như: Đất trồng, đường giao thông, mặt nước, rừng cây... Thông thường, các mã phân loại đại diện cho kiểu đối tượng được thu nhận trong tín hiệu phản hồi. Phân loại đám mây điểm là bước quan trọng trong quá trình trích xuất thông tin của các lớp như tòa nhà, thực vật, giao thông và mặt nước. Thuật toán phân loại sử dụng K-means sẽ lựa chọn các điểm mẫu trong mẫu ngẫu nhiên từ toàn bộ đám mây điểm. Phương pháp phân loại được thể hiện qua sơ đồ hình 3.



Hình 3. Quy trình phân loại đám mây điểm LiDAR sử dụng K-means.

Sử dụng thuật toán K-means trong phân loại đám mây điểm LiDAR được thực hiện như sau:

Để thử nghiệm thuật toán K-means trong phân loại đám mây điểm, chúng tôi đã tiến hành thực nghiệm với bộ dữ liệu được đo tại thành phố Vinh, tỉnh Nghệ An với

số điểm demo là 538 điểm, mỗi điểm gồm giá trị (x, y, z), trong đó giá trị z được lấy là thuộc tính để tiến hành phân loại.

Với số cụm được lựa chọn ngẫu nhiên, qua mỗi lần thử nghiệm kết quả cho được là khác nhau. Đầu tiên, với $k = 2$, qua 2 lần lặp sẽ phân chia được hai cụm với kết quả như hình 4.

	Cluster	
	1	2
V4	1146	12198

Cluster	1	460,000
	2	78,000
Valid		538,000
Missing		1763,000

Cụm khởi tạo Kết quả phân cụm
Hình 4. Phân lớp với số cụm khởi tạo $k = 2$.

Khi số cụm được khởi tạo với giá trị $k = 10$, kết quả được thể hiện trong hình 5, 6 và 7.

	Cluster									
	1	2	3	4	5	6	7	8	9	10
V4	11945	11580	12198	11696	1146	1496	3099	1796	2099	11384

Hình 5. Tâm của các cụm khởi tạo.

	Cluster									
	1	2	3	4	5	6	7	8	9	10
V4	11982	11587	12185	11691	1193	1547	3102	1854	2088	11390

Hình 6. Sự thay đổi của tâm cụm khi phân loại.

Cluster	1	14,000
	2	10,000
	3	20,000
	4	19,000
	5	407,000
	6	15,000
	7	23,000
	8	8,000
	9	7,000
	10	15,000
Valid		538,000
Missing		1763,000

Hình 7. Các điểm trong các nhóm sau phân loại với $k = 10$.

Khi số cụm khởi tạo được tăng lên là 20, số lần lặp là 6, trọng tâm của các cụm mới được tạo và số điểm được gán vào các cụm cũng thay đổi, kết quả được thể hiện trong hình 8.

Cluster	1	10,000
	2	2,000
	3	1,000
	4	2,000
	5	6,000
	6	19,000
	7	2,000
	8	230,000
	9	9,000
	10	2,000
	11	1,000
	12	20,000
	13	3,000
	14	10,000
	15	15,000
	16	7,000
	17	28,000
	18	19,000
	19	147,000
	20	5,000
Valid		538,000
Missing		1763,000

Hình 8. Số điểm được gán vào cụm khi số cụm khởi tạo thay đổi.

Và khi k tăng lên là 30 thì kết quả của thuật toán với số cụm được tính toán lại trọng tâm được thể hiện trong hình 9.

Cluster	1	37,000
	2	3,000
	3	2,000
	4	7,000
	5	79,000
	6	2,000
	7	7,000
	8	3,000
	9	2,000
	10	2,000
	11	13,000
	12	2,000
	13	59,000
	14	2,000
	15	1,000
	16	2,000
	17	6,000
	18	2,000
	19	14,000
	20	1,000
	21	3,000
	22	1,000
	23	117,000
	24	10,000
	25	3,000
	26	2,000
	27	2,000
	28	1,000
	29	11,000
	30	2,000
Valid		538,000
Missing		1763,000

Hình 9. Kết quả phân cụm với $k = 30$.

Qua 4 thử nghiệm với lựa chọn số cụm khởi tạo khác nhau, có thể thấy thuật toán K-means hoạt động khá ổn

định khi giá trị missing của thuật toán không thay đổi khi k tăng dần. Do đó, kết quả phân loại là đáng tin cậy và có thể ứng dụng trong bài toán phân loại đám mây điểm LiDAR.

Kết luận

Thuật toán K-means là thuật toán điển hình trong bài toán phân cụm dữ liệu, là giải thuật dễ hiểu và dễ cài đặt. Với dữ liệu thử nghiệm, khi số cụm tăng dần, giá trị missing của thuật toán không thay đổi, vì thế thuật toán K-means hoàn toàn phù hợp với bài toán phân loại dữ liệu đám mây điểm LiDAR. Tuy nhiên, thử nghiệm mới chỉ sử dụng một thuộc tính (độ cao của dữ liệu điểm) trong phân cụm, việc giải quyết bài toán có nhiều thuộc tính sẽ là hướng phát triển tiếp theo của nghiên cứu này.

TÀI LIỆU THAM KHẢO

- [1] Trần Đình Trí (2013), *Công nghệ LiDAR*, Đại học Mở - Địa chất.
- [2] D. Albashish (2011), "Detection and classification of leaf diseases using K-means-based segmentation and Neural Network-based classification", *Information Technology Journal*, **10**, pp.267-275.
- [3] Balasubramanian (2012), "Image classification through

intergrated K-means algorithms", *International Journal of Computer Science Issues*, **9(2)**, pp.518-524.

[4] Kun Zhang, Weihong Bi, Xiaoming Zhang, Xinghu Fu (2015), "A new K-means clustering algorithm for point cloud", *International Journal of Hybrid Information Technology*, **8**, pp.157-170.

[5] Trần Đình Luật (2015), "Khả năng ứng dụng công nghệ LiDAR xây dựng mô hình số địa hình vùng bãi bồi cửa sông ven biển trong điều kiện Việt Nam", *Tạp chí Tài nguyên và Môi trường*, **5**, tr.830-833.

[6] Lương Chính Kế (2010), *Thành lập DEM/DTM bằng công nghệ, Viện Đo ảnh và bản đồ, Đại học Bách khoa Vacsava*.

[7] Trần Đức Phú (2010), *Ứng dụng công nghệ LiDAR trong mô hình hóa lũ*, Trường Đại học Hàng hải.

[8] K. Kosi (2012), "Methods of data center classification", *Acta Polytechnica Hungarica*, **9(5)**, pp.127-137.

[9] CEE6150 (2015), *Unsupervised Classification (Cluster analysis)*, US: s.n.

[10] E. Sigova (2015), *A semi supervised approach to dialogue act classification using K-means +HMM*, KTH of Computer Science and communication.

[11] Balasubramanian (2012), "Image classification through intergrated K-means algorithms", *International Journal of Computer Science Issues*, **9(2)**, pp.518-524.