Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất
Trang điện tử: http://tapchi.humg.edu.vn

Thuật toán K-means và k-NN trong phân loại đám mây điểm Lidar

Nguyễn Thị Hữu Phương *

Khoa Công nghệ thông tin, Trường Đại học Mỏ - Địa chất, Việt Nam

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 20/6/2017
Chấp nhận 20/7/2017
Đăng online30/10/2017

Từ khóa:
Lidar
K-means
k-NN
Phân loại

TÓM TẮT

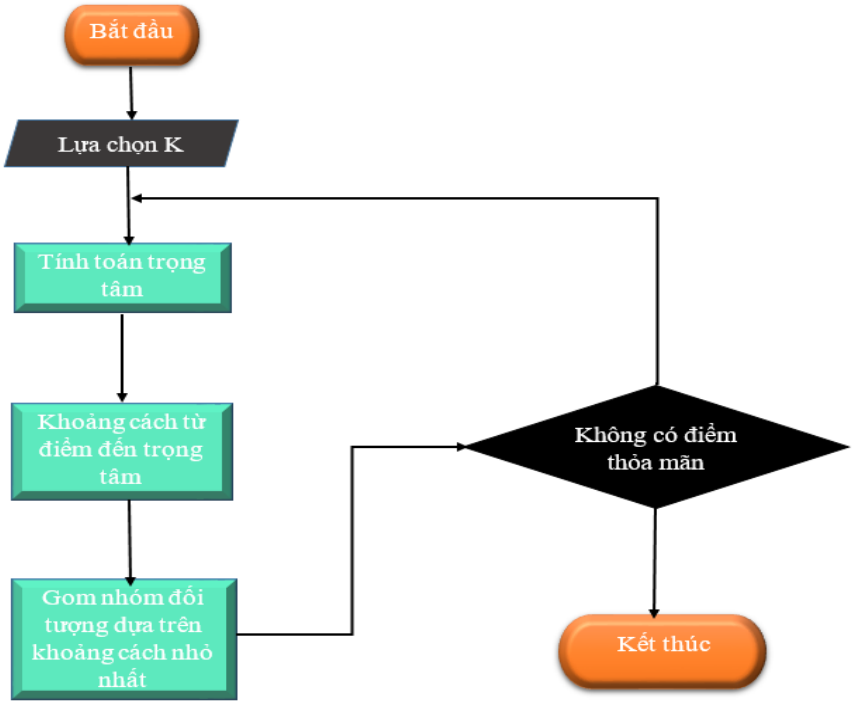
Thuật toán K-means và k-NN (k – Nearest Neighbor) là hai thuật toán rất phổ biến trong khai phá dữ liệu. K-means là thuật toán phân cụm thuộc nhóm phân loại không giám sát, với ý tưởng nhóm đối tượng vào k cụm với trọng tâm của mỗi cụm thay đổi sau mỗi lần lặp. k-NN là thuật toán phân loại thuộc nhóm phân loại có giám sát, thuật toán sẽ tính toán khoảng cách từ đối tượng đến tâm các cụm, tìm giá trị khoảng cách nhỏ nhất và gán đối tượng vào lớp tương ứng. Bài báo tập trung vào nghiên cứu ứng dụng của hai thuật toán K-means và k-NN vào bài toán phân loại đám mây điểm Lidar - dữ liệu viễn thám có độ chính cao và số lượng điểm tương đối lớn. Với bộ dữ liệu thử nghiệm là 485 điểm được đo tại Nghệ An, kết quả phân loại dựa trên giá trị độ cao của điểm cho thấy giá trị lỗi khi phân loại với hai thuật toán vẫn còn chiếm tỉ lệ khá cao với thuật toán K-means (31,5%) và thuật toán k-NN là 48.4%.

1. Mở đầu

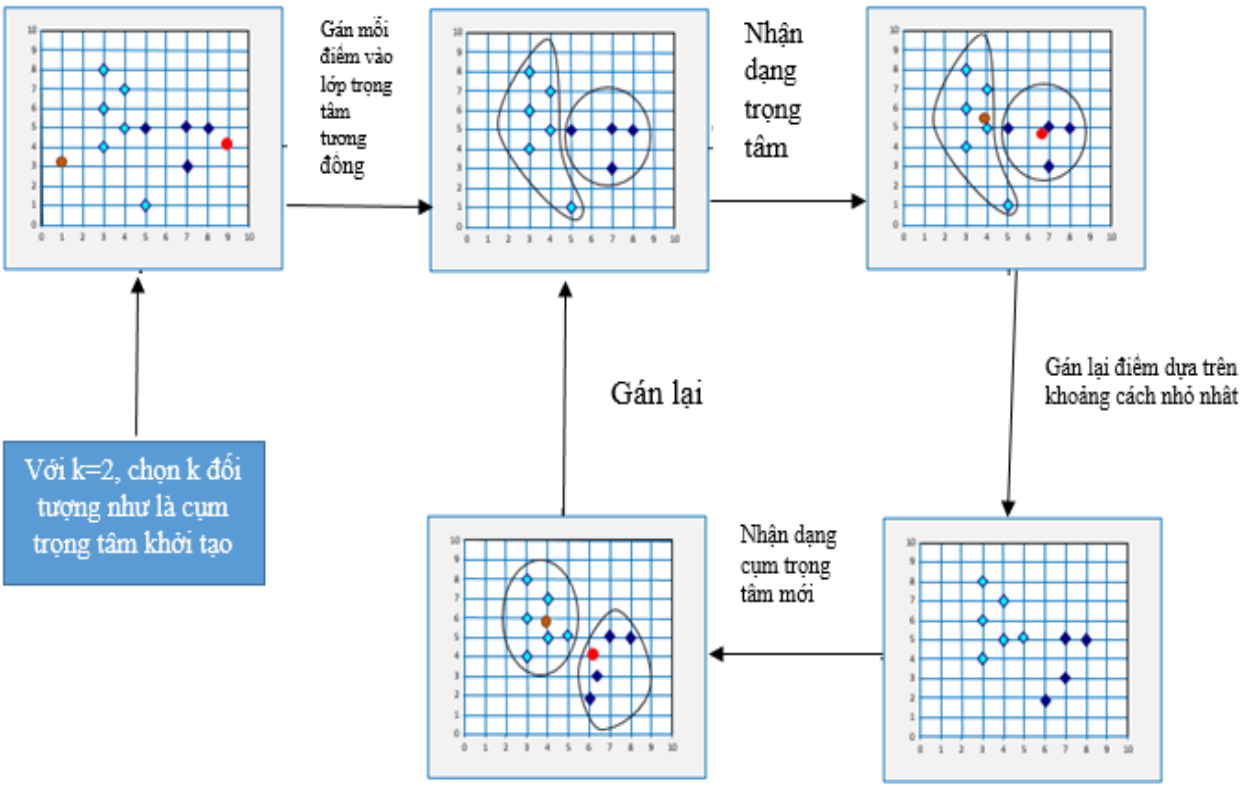
Lidar là công nghệ viễn thám mới, chủ động, sử dụng các loại tia laser để khảo sát đối tượng từ xa. Dữ liệu thu được của hệ thống là tập hợp đám mây điểm phản xạ 3 chiều của tia laser từ đối tượng được khảo sát. Hiện nay, công nghệ Lidar đang được ứng dụng rộng rãi trong việc: khảo sát địa hình và lập 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 (DEM), mô hình số địa hình (DTM) hay mô hình số bề mặt (DSM), phải tiến hành

*Tác giả liên hệ
E-mail: nguyenphuong85.nb@gmail.com



Hình 1. Mô tả thuật toán K-means.



Hình 2. Ví dụ thuật toán K-means.

phân loại điểm trong đám mây điểm đó (Trần Đình Trí, 2013).

Hiện nay, có nhiều thuật toán phân loại đám mây 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, thì 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 (Trần Đình Trí, 2013).

Trên thế giới, việc phân loại dữ liệu Lidar để từ đó trích xuất ra được các đối tượng phục vụ trong công tác xây dựng bản đồ và nhiều lĩnh vực khác của đời sống xã hội đã khá phổ biến. Trong các nghiên cứu (Borja Rodriguez – Cuenca et al., 2015) (K.Rumkis et al., 2014) (Kun Zhang et al., 2015) (Yu-chuan Chang et al., 2008) (Zhuqiang Li at el., 2016) 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 (Trần Đình Luật, và nnk, 2015) 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 (Lương Chính Kế, 2005) và Trần Đức Phú (Trần Đức Phú, 2010) 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.

Khai phá dữ liệu với các kĩ thuật và thuật toán phân loại đang dần được sử dụng tương đối phổ biến khi những tri thức con người cần khai phá từ những dữ liệu thu được trở nên cần thiết. Với những hạn chế từ cách thực hiện và phương pháp phân loại đám mây điểm Lidar tại Việt Nam, tôi đã tiến hành nghiên cứu sử dụng thuật toán K-means và k-NN 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.

2. Thuật toán K-means trong phân loại

Thuật toán K-means là tìm phương pháp phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các 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 hình 2.

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 như sau

Input: Dữ liệu LiDAR
Output: Bộ dữ liệu sau phân loại
Procedure
1. Initial: Chọn số cụm được khởi tạo, số lớp cần phân loại n
   If (k > n), thuật toán kết thúc
   ElseIf (k ≤ n), thì chọn k ngẫu nhiên, tính toán trọng tâm của các cụm vừa tạo
2. While(1)
   Tính toán khoảng cách của các điểm đến trọng tâm của cụm $d_0(x_i,k)$
   Tìm các nhóm điểm thỏa mãn $d_i = d_{min}(x_i,k)$, G
   If ($d_{i+1} \neq d_i$)
      Cập nhật các cụm mới
      Tính toán lại trong tâm của các cụm mới
   Else
      If (k = n) hoặc $d_{i+1} = d_i$
   Kết thúc lặp
3. Với k thu được sau quá trình lặp, gán tên thuộc tính với k_i
4. Kết thúc thuật toán

Hình 3. Pseudo code của thuật toán

(Alex Berson et al., 2005):

Bước 1: chọn K cụm trọng tâm khởi tạo, z_1, z_2, z_3, …, z_n, với $0 < K \leq n$

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 $j = 1, 2, 3, …, k; i \neq j$; $S_i(n)$ là bộ mẫu của trọng tâm $z_i(n)$, trong đó n 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:

$$z_i(n+1) = \frac{1}{c}\sum_{x \in S_i(n)} x$$

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

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

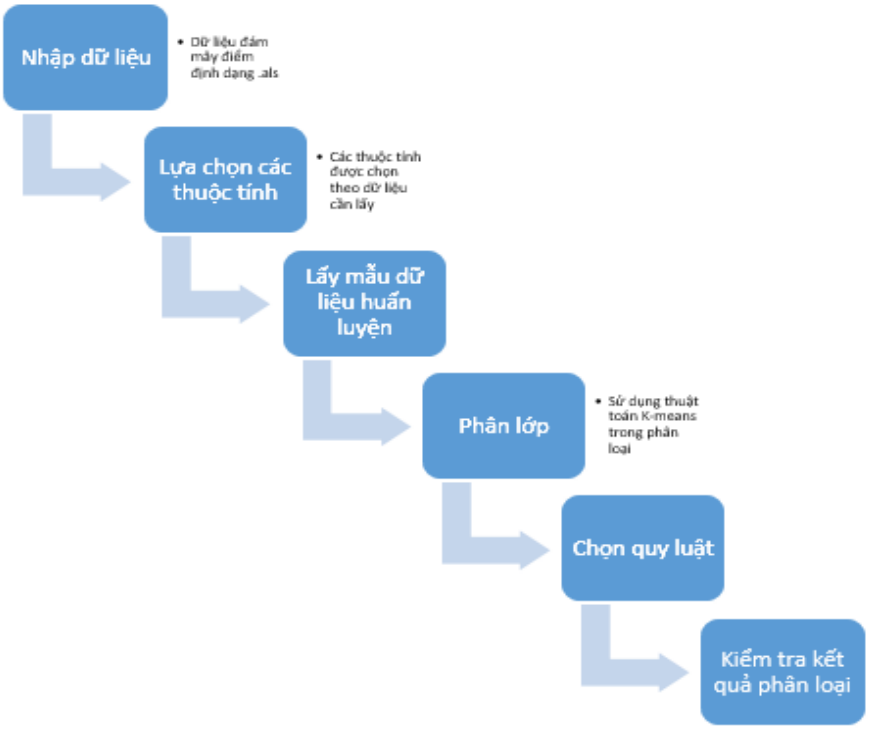
Trong thuật toán K-means việc chọn được giá trị k sẽ có thể giúp tăng tốc được thuật toán, tối ưu và cải tiến thuật toán tốt hơn. Có nhiều phương pháp để có thể lựa chọn được giá trị k như sử dụng ý kiến của chuyên gia, thử mô hình với các giá trị của k và từ đó chọn k tốt nhất hay sử dụng kỹ thuật CV (Cross - Validation), ….

3. Thuật toán k-NN trong phân loại

k-NN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần phân lớp và tất cả các đối tượng trong dữ liệu huấn luyện. Phương pháp k-NN sẽ tìm K điểm trong bộ dữ liệu huấn luyện mà gần với điểm cần phân lớp nhất. Sau đó, điểm này sẽ được gán vào lớp mà đa số láng giềng của nó thuộc về. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Kĩ thuật phân loại sử dụng kNN để gán lớp cho đối tượng Z_i được thực hiện như sau (M.Ranageri, 2010):

o Cho X là tập dữ liệu huấn luyện với nhãn lớp là c, tập X được kí hiệu $X = \{(x, c)\}$, Z là bộ dữ liệu cần gán lớp, r là khoảng cách hình học, C là tập nhãn của lớp.

o Tính toán khoảng cách giữa điểm Z_i đến X_i kí hiệu là $R(Z_i, X_i)$



Hình 4. Quy trình phân loại đám mây điểm Lidar của thuật toán K-means

o Xác định khoảng cách ngắn nhất R_{min}, và lựa chọn tập $Uk(z)$ là tập của k mẫu huấn luyện gần nhất với Z_i

o Gán Z_i vào lớp có nhãn là C mà có chứa những điểm hàng xóm gần với Zi nhất

Đặc trưng của kĩ thuật kNN là xác định một số mẫu huấn luyện hoặc nguyên mẫu của nó, đây là phương pháp phân loại có độ chính xác dựa hoàn toàn vào khoảng cách. Do đó, nó là phương pháp phù hợp với ứng dụng dự đoán kết quả. Quá trình huấn luyện của phân lớp với k-NN là tương đối đơn giản, nhưng quá trình kiểm tra của k-NN sẽ chậm hơn.

Với k-NN việc tính khoảng cách từ điểm kiểm tra đến dữ liệu huấn luyện sẽ quyết định độ chính xác của lớp mà nó thuộc về, do đó quá trình này vô cùng quan trọng. Công thức để tính toán khoảng cách hay được dùng trong k-NN là:

Cho hai điểm $X = (x1, x2, ..., x2)$, $Y = (y1, y2,..., yn)$ trong không gian R_2. Khoảng cách từ một điểm p đến hai điểm X, Y được định nghĩa:

$$D_p = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \quad (1)$$

Trong đó:

Nếu $p = 1$ khoảng cách này là khoảng cách Manhattan

Nếu $p = 2$ đây là khoảng cách Euclide

Nếu $p = \infty$, khoảng cách vô cùng được định nghĩa theo công thức

$D_i = max\{|x_i - y_i|\}$

4. Phân loại đám mây điểm Lidar với K-means và k-NN

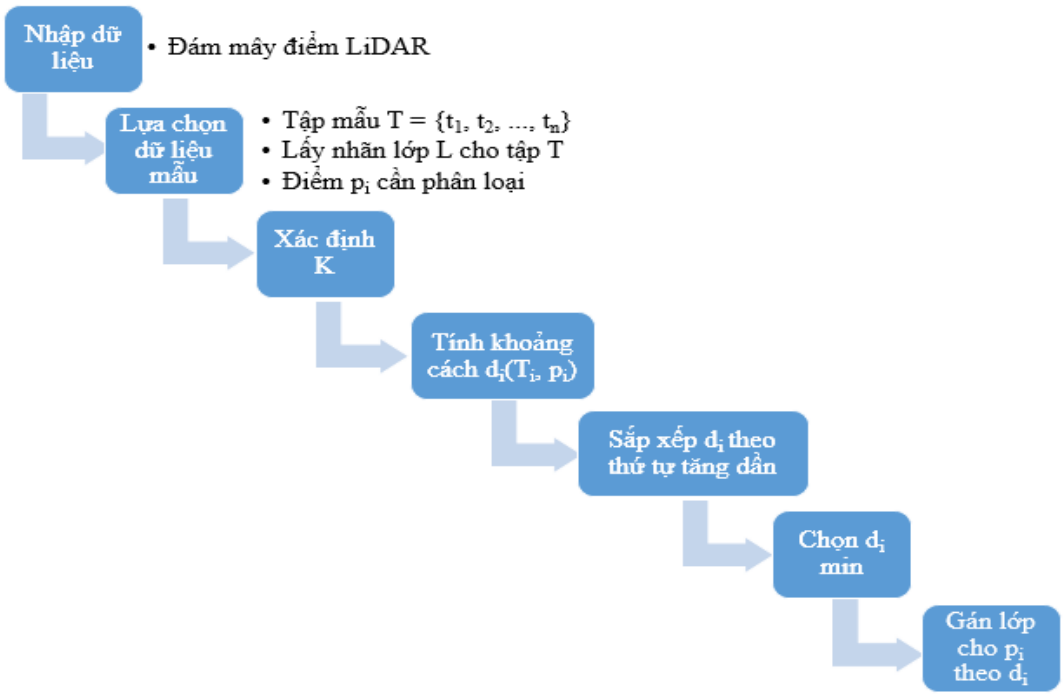
4.1 Đặc trưng đám mây điểm Lidar

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 (Trần Đình Trí, 2013).

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.



Hình 5. Quy trình thực hiện của thuật toán.

• Cường độ xung phản xạ (Intensity): khi tia laser phản xạ trở lại nó 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 các thuộc tính của đám mây điểm Lidar như: số phản hồi, góc máy bay, thời gian GPS, góc quét, hướng quét …

4.2 Thuật toán K-means trong phân loại đám mây điểm Lidar

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 trống, thực vật cao, thực vật thấp, và nước .... 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 4.

4.3 Phân loại đám mây điểm Lidar với thuật toán k-NN

Thuật toán k-NN được sử dụng để phân loại đám mây điểm Lidar được thể hiện như sau:

Input: đám mây điểm Lidar *P*, tập mẫu *T* đã được phân loại, $p_i \in P$ cần gán lớp, *L* là nhãn lớp của của *T*

Output: p_i đã được gán lớp

Procedure

For *i =1* to *n* do

Tính khoảng cách d(T_i, p_i)

Sắp xếp khoảng cách theo thứ tự tăng dần

End for

Lấy k giá trị đầu tiên trong tập giá trị ngắn nhất

Tìm k điểm tương ứng với k giá trị

If $k_i > k_j$? i ≠ j then

Gán điểm p_i vào lớp i

End.

4.4 Kết quả thử nghiệm

Để có thể đánh giá được khả năng phân loại đám mây điểm Lidar của hai thuật toán K-means và k-NN, tác giả đã thử nghiệm phân loại

Initial Cluster Centers

	Cluster				
	1	2	3	4	5
V4	17024	12766	4572	7159	10184

Số cụm khởi tạo

Iteration History[a]

Iteration	Change in Cluster Centers				
	1	2	3	4	5
1	543,000	234,500	950,489	770,335	515,778
2	,000	,000	123,320	97,904	267,222
3	,000	,000	76,174	114,969	118,750
4	,000	,000	38,064	75,833	,000
5	,000	,000	12,973	35,579	118,750
6	,000	,000	9,113	28,007	135,800
7	,000	,000	4,621	8,552	,000
8	,000	,000	2,338	4,279	,000
9	,000	,000	2,339	4,311	,000
10	,000	,000	2,322	4,379	,000

Số lần lặp của thuật toán

Final Cluster Centers

	Cluster				
	1	2	3	4	5
V4	16481	12532	5796	6767	9537

Tâm mới của các cụm

Number of Cases in each Cluster

Cluster	1	2,000
	2	2,000
	3	208,000
	4	110,000
	5	10,000
Valid		332,000
Missing		153,000

Số lượng điểm sau phân loại của mỗi nhóm

Hình 6. Kết quả phân loại với k = 5.

Initial Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
V4	9029	10518	6038	7537	12766	4572	17024

Tâm khởi tạo của các cụm

Iteration History[a]

Iteration	Change in Cluster Centers						
	1	2	3	4	5	6	7
1	138,571	119,667	66,508	349,500	234,500	353,091	543,000
2	,000	,000	65,548	236,682	,000	156,690	,000
3	,000	,000	59,619	103,407	,000	26,542	,000
4	140,571	,000	25,862	47,217	,000	13,384	,000
5	116,778	,000	11,948	31,831	,000	,000	,000
6	103,822	,000	2,297	14,873	,000	,000	,000
7	92,127	,000	,000	10,292	,000	,000	,000
8	79,523	,000	2,250	14,040	,000	,000	,000
9	182,850	339,733	4,587	3,845	,000	13,158	,000
10	324,438	101,600	2,341	37,610	,000	13,846	,000

Lịch sử lặp của thuật toán

Final Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
V4	7913	9748	5923	6672	12532	5053	16481

Tâm mới của cụm

Number of Cases in each Cluster

Cluster	1	12,000
	2	8,000
	3	179,000
	4	99,000
	5	2,000
	6	30,000
	7	2,000
Valid		332,000
Missing		153,000

Số lượng điểm được phân vào mỗi nhóm sau khi phân loại

Hình 7. Kết quả phân loại thuật toán K-means với k = 7

Case Processing Summary

		N	Percent
Sample	Training	235	94,0%
	Holdout	15	6,0%
Valid		250	100,0%
Excluded		235	
Total		485	

Hình 8. Kết quả xử lý.

với bộ dữ liệu được đo tại Nghệ An, với 485 điểm thử nghiệm, mỗi điểm được thể hiện với 3 thuộc tính (x, y, z), trong đó thuộc tính được sử dụng để phân loại là z (giá trị độ cao của điểm). Hai thuật toán được chạy với phần mềm SPSS 20 của IBM.

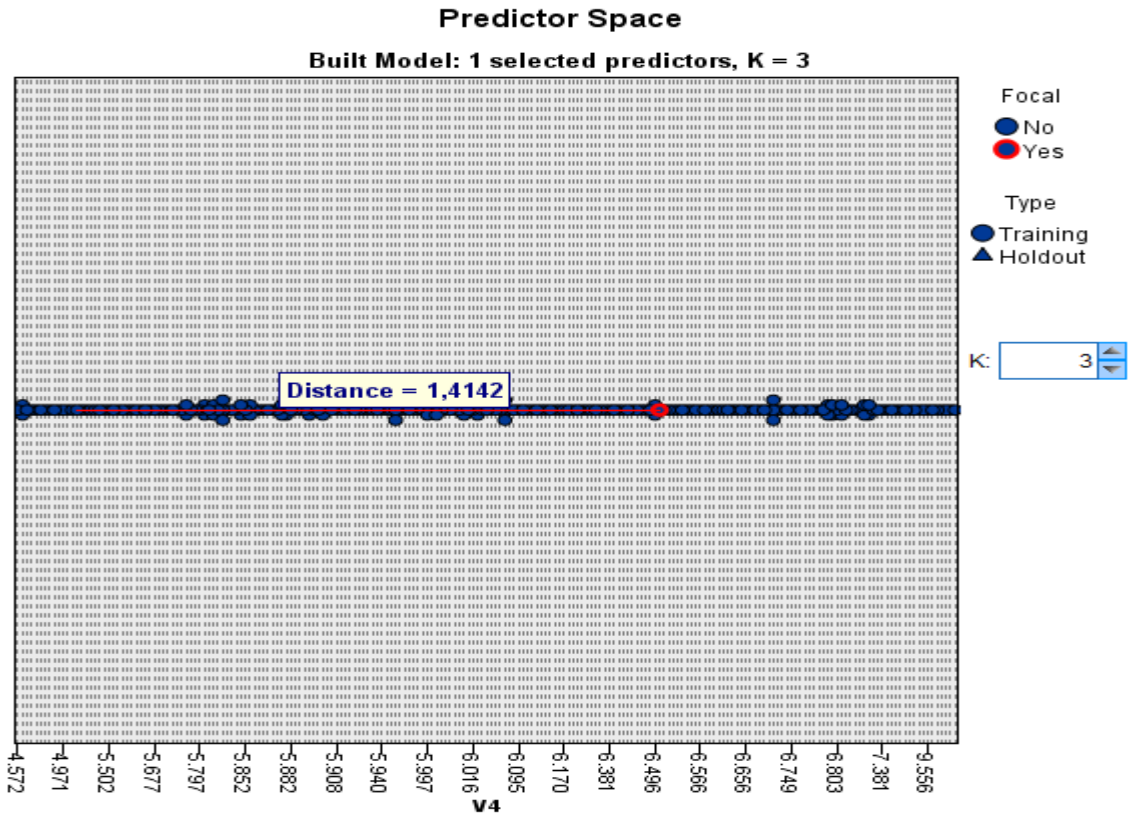
• Với thuật toán K-means, việc lựa chọn số lượng cụm thích hợp cho một bộ dữ liệu nhất định trong thuật toán K-Means sẽ quyết định đến cụm trong quá trình phân cụm. Đây là một quá trình khó khăn vì kết quả của quá trình phân cụm do người sử dụng quyết định. Sự lựa chọn chính xác K thường không rõ ràng, do sự phân bố và quy mô của các điểm trong bộ dữ liệu và độ phân giải của người dùng. Để có thể tìm được K phù hợp với bộ dữ liệu, thông thường người dùng sẽ chạy thuật toán K-Means với các giá trị K khác nhau và so sánh kết quả để lựa chọn được K phù hợp. Để thử nghiệm với bộ dữ liệu, tác giả đưa ra hai giá trị K để tiến hành phân loại là K = 5 và K = 7.

- Với lựa chọn k = 5, kết quả phân loại được thể hiện như sau:

- Với k = 5, qua 10 lần lặp thuật toán đã phân chia được 332 điểm vào 5 cụm, có 153 giá trị lỗi (điểm không được phân về cụm nào). Với trọng tâm của 5 cụm được khởi tạo như trong hình 6 – số cụm khởi tạo, với 10 lần lặp, trọng tâm của 5 cụm được tính toán lại như trong hình 6 – tâm mới của cụm. Kết quả trong tổng số 332 điểm có 2 điểm thuộc về cụm 1, 2 điểm cụm 2, 208 điểm cụm 3, 110 điểm cụm 4 và 10 điểm cụm 5.

- Với k = 7

Khi tăng số cụm lên là 7, giá trị điểm không được gán vào cụm nào không thay đổi là 153 điểm, trọng tâm của cụm được lựa chọn như trong Hình 7 – số cụm khởi tạo, qua số lần lặp là 10, trọng tâm của cụm được tính toán lại như trong Hình 7 – tâm mới của cụm. Kết quả có 12 điểm được gán vào cụm 1, 8 điểm được gán vào cụm 2, 179 điểm cụm 3, 99 điểm cụm 4, 2 điểm cụm 5, 30 điểm cụm 6 và 2 điểm cụm 7.



Hình 9. Lựa chọn điểm cần phân loại

k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distances		
	1	2	3	1	2	3
230	332	331	330	1,414	1,414	1,414

Hình 10. Điểm gần nhất và khoảng cách gần nhất cho điểm chọn.

• Trong khi đó, thuật toán k-NN là thuật toán phân loại không sử dụng kết quả học từ bộ dữ liệu huấn luyện mà hoàn toàn phụ thuộc vào số điểm lân cận với điểm cần khảo sát. Do đó, kết quả phân cụm phụ thuộc vào giá trị K. Để đánh giá được ảnh hưởng K tới quá trình phân loại, tác giả lựa chọn giá trị K = 3 và K = 1.

- Với giá trị K = 3:

Trong tổng số 485 điểm được đưa vào phân loại, chọn được 250 giá trị hợp lệ, loại bỏ 235 giá trị. Trong số 250 điểm chấp nhận, có 235 điểm được đưa vào huấn luyện, 15 điểm được sử dụng trong quá trình kiểm tra, đánh giá kết quả.

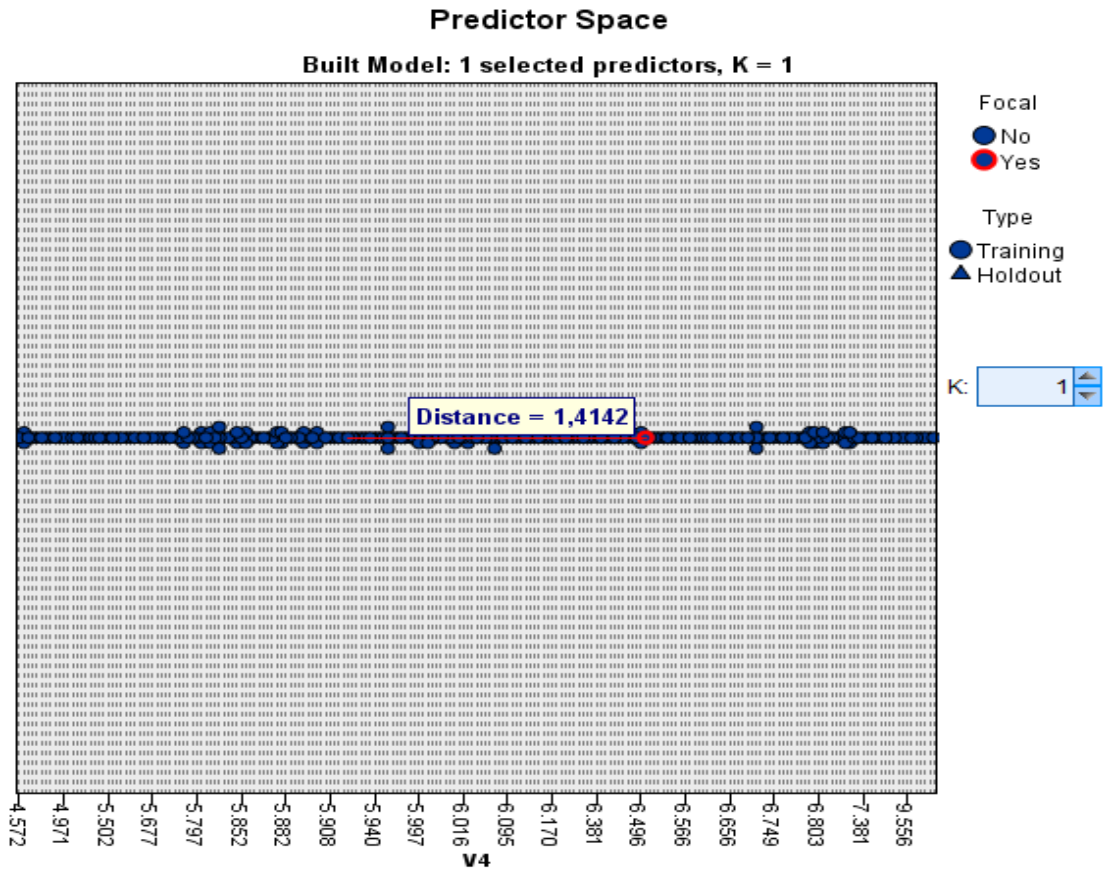
Tiến hành lựa chọn điểm 230 để đưa vào dự đoán (điểm đánh dấu đỏ trong hình 9), với K = 3, có 3 điểm gần nhất với điểm 230 là 332, 331 và 330 với khoảng cách gần nhất cùng là 1,414.

Khi thay đổi giá trị k = 1, kết quả phân loại thay đổi như sau:

Khi K = 1, lúc này điểm gần nhất so với điểm 230 là điểm 332 và khoảng cách gần nhất là 1,414.

4.5. Đánh giá kết quả

Sau khi thử nghiệm với bộ dữ liệu trên hai thuật toán K-means và k-NN, nhận thấy đây là hai thuật toán phân loại dựa hoàn toàn vào khoảng cách. Các điểm được gán vào cụm (lớp) phụ thuộc vào khoảng cách của nó tới tâm cụm



Hình 11. Lựa chọn K và điểm cần phân lớp.

k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors	Nearest Distances
	1	1
230	332	1,414

Hình 12. Điểm lựa chọn.

(với K-means) hay tới điểm hàng xóm gần nhất với nó (k-NN). Thuật toán K-means có sự thay đổi rõ ràng nhất về sự thay đổi của phân bố các điểm trong cụm và trọng tâm được tính toán trong mỗi cụm khi có sự thay đổi của số cụm khởi tạo. Sự thay đổi này được thể hiện trong Bảng 1.

Tuy nhiên, với cả 2 giá trị k số lượng điểm lỗi chiếm tỉ lệ khá lớn 153/485 điểm (31.5%). Do vậy, K-means là thuật toán có độ chính xác phụ thuộc vào quá trình tính toán trọng tâm của cụm qua mỗi lần lặp.

Trong khi lựa chọn cụm để gán trong thuật toán k-NN phụ thuộc vào số điểm lân cận mà người dùng lựa chọn. Khi chọn giá trị K = 3 có 3 điểm được lựa chọn là 332, 331, 330, khi lựa chọn giá trị K = 1 lúc này có điểm 332 được gán trọng số cao hơn so với hai điểm còn được lựa chọn là điểm hàng xóm gần nhất với điểm 230. Có thể nhận thấy, kết quả của thuật toán thay đổi khi lựa chọn số điểm lân cận, quá trình lựa chọn lớp nào là lớp mà điểm dự báo thuộc về phụ thuộc hoàn toàn vào khoảng cách ngắn nhất và quá trình lựa chọn điểm.

Qua thử nghiệm với hai thuật toán K-means và k-NN, tác giả đưa ra Bảng 2:

Bảng 1. Trọng tâm và số điểm thay đổi khi thay đổi số cụm khởi tạo trong K-mea.

	Cụm	Trọng tâm	Số điểm		Cụm	Trọng tâm	Số điểm
k = 5	1	16,481	2	k = 7	1	7,913	12
	2	12,532	2		2	9,748	8
	3	5,796	208		3	5,923	179
	4	6,767	110		4	6,672	99
	5	9,537	10		5	12,532	2
					6	5,053	30
					7	16,481	2

Bảng 2. So sánh kết quả thử nghiệm K-means và k-NN

STT	Tiêu chí	K-means	k-NN
1	Thuộc tính phân loại	Độ cao điểm	Độ cao điểm
2	Thời gian chạy thuật toán (giây)	- K = 5: 2s - K = 7: 5s	- K = 3: 9s - K = 1: 9s
3	Giá trị lỗi	153	235
4	Điều kiện phân loại	Khoảng cách giữa tâm của cụm đến điểm cần phân loại	Khoảng cách từ điểm cần phân loại đến các điểm gần nhất trong bộ huấn luyện
5	Giá trị ảnh hưởng độ chính xác	Số cụm khởi tạo	Khoảng cách từ điểm dự báo đến điểm gần nhất
6	Độ chính xác	68.5%	51.6%

5. Kết luận

Thuật toán K-means và k-NN là hai thuật toán phổ biến của bài toán phân loại dữ liệu, hai thuật toán này dễ cài đặt và thử nghiệm. Với bộ dữ liệu thử nghiệm 485 điểm được đo tại Nghệ An có thể thấy, thuật toán K-means phân loại có độ chính xác cao hơn với 68.5% điểm chính xác, trong khi thuật toán k-NN là 51.6%. Nguyên nhân là do thuật toán k-NN lựa chọn lớp để gán điểm không dựa trên quá trình học từ bộ dữ liệu huấn luyện, mà chỉ khi nào có yêu cầu phân loại, thuật toán mới tiến hành tính toán khoảng cách từ điểm dự báo đến số điểm lân cận mà người dùng lựa chọn. Chính vì thế, để có thể tăng độ chính xác của bài toán, cần phải cải tiến hai thuật toán này cho phù hợp với dữ liệu đám mây điểm Lidar.

Tài liệu tham khảo

Alex Berson, 2005. An overview of data mining technique. Building DM Applications for CRM. IEEE.

Borja Rodriguez - Cuenca, Silverio Garcia Cortes, Celestino Ordonez, Maria C.Alonso. (2015). Automatic detection and classification of pole-like objects in urban point cloud data using an anomaly detection algorithm. Remote Sensing, 7, 12680-12703.

Jiawei Han, Micheline Kamber., 2000. Data mining concept and techniques. Morgan Kaufman.

Kapourani, C. A., 2016. K-means clustering and kNN classification. IEEE.

Kun Zhang, Weihong Bi, Xiaoming Zhang, Xinghu Fu, Kunpeng Zhu, Li Zhu., 2015. A new kmeans clustering algorithm for point cloud. International Journal of Hybrid Information Technology, 8(9), 157-170.

Lương Chính Kế, 2005. Thành lập DEM/DTM DSM bằng công nghệ Lidar. Tạp chí Tài nguyên và Môi trường.

Mansi Gera, Shivani Goel., 2015. Data mining - Techniques, methods and algorithms: a review on tool and their validity. International of Computer Applications , 22-29.

Ranageri, M. B., 2010. Data mining techniques and applications. Indian Journal of Computer Scienes and Engineering, (pp. 301-305). India.

Rumkis, K., Lipnickas, A., Sinkevicius, S., 2014. Classification of 3D Point cloud using numerical surface signatures on interest point. Elektronika IR elektrotechnika, 20(6), 8-11.

Seyed Hossein Hosseini Nourzad, Anu Pradhan., 2012. Binary and multi-class classification of fused Lidar - Imagery data using an ensemble method. Construction Research Congress .

Slota, M., 2014. Advanced processing techniques and classification of full-wave form ALS data. Geomatics and Environmental engineering, 8(2), 85-95.

Suresh K.Lodha, Darren M.Fitzpatrick, David P.Helmbold., 2007. Aerial Lidar data classification using Expectation - Maximization. IEEE.

Trần Đình Luật, Nguyễn Thị Kim Dung, Lưu Thị Thu Thủy, Trần Hồng Hạnh, 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, 1, 24-28.

Trần Đình Trí, 2013. Bài giảng Công nghệ Lidar. Hà Nội: Nhà xuất bản Giao thông vận tải.

Trần Đức Phú., 2010. Ứng dụng công nghệ Lidar trong mô hình hóa lũ. Tạp chí Khoa học công nghệ và hàng hải, 23, 54-58.

Yu-chuan Chang, Ayman F.habib, Dong Cheon Lee, Jae Hong Yom, 2008. Automatic classification of Lidar data into Ground and non-Ground points. The International Archives of the Photogrammetry, RS and Spatial Information Sciences, XXXVII(B4), 457-462.

Zhuqiang Li, Li Qiang Zhang, Xiaohua Tong, 2016. A three step approach for TLS point cloud classification . IEEE.

ABSTRACT

K-NN and K-means algorithms in Lidar point cloud classification

Phuong Thi Huu Nguyen *,

[1] *Hanoi University of Mining and Geology, Bac Tu Liem District, Hanoi City, Vietnam.*

The k-NN (k - Nearest Neighbor) algorithm is common algorithms for data mining. K-means is a clustering algorithm belonging to the unsupervised classification, with the idea of grouping objects into k clusters with the focus of each clustered change after each iteration. k-NN is the supervised classification, which calculates the distance from the object to the center of the cluster, finds the smallest distance value, and assigns the object to the corresponding class. This article focuses on the applied research of two K-means and k-NN algorithms into the Lidar cloud point classification - high accuracy remote sensing data and large number of points. With a test data set of 485 points measured in Nghe An, the classification result based on the elevation point value indicates that the error value classified with two algorithms still accounts for the high K -means (31.5%) and the k-NN algorithm is 48.4%.

Keywords: K-means, k-NN, Lidar, classification.