Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00039

KHAI PHÁ DỮ LIỆU LIDAR TRONG NGHIÊN CỨU CÁC ĐỐI TƯỢNG TRÊN BỀ MẶT ĐỊA HÌNH

Nguyễn Thị Hữu Phương[1], Đặng Văn Đức[2], Nguyễn Trường Xuân[1]

[1] Trường Đại học Mỏ - Địa chất

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

nguyenphuong85.nb@gmail.com, dvduc@ioit.ac.vn, nguyentruongxuan@humg.edu.vn

TÓM TẮT: *LiDAR từ khi ra đời vào thập niên 90 của thế kỷ XX đến nay đã được áp dụng vào nhiều ngành và nhiều lĩnh vực của đời sống xã hội. Với khả năng thu nhận dữ liệu về khu vực quét rộng lớn, đo được ban đêm và không bị ảnh hưởng bởi thời tiết, thêm vào đó, dữ liệu thu nhận được từ LiDAR có độ chính xác cao. So với các kỹ thuật của trắc địa truyền thống, LiDAR có những ưu điểm và ưu thế nổi bật trong nghiên cứu đối tượng trên bề mặt địa hình. Tuy nhiên, dữ liệu thu được từ hệ thống LiDAR là tương đối lớn, với khoảng từ 4000 - 5000 điểm trên một km^2 bao gồm thông tin tọa độ, thời gian bay quét, độ cao,…. Để sử dụng có hiệu quả dữ liệu LiDAR trong nghiên cứu các đối tượng trên bề mặt địa hình cần phải có các kĩ thuật khai thác nhằm tìm kiếm những thông tin có ích. Bài báo này tập trung vào trình bày các kĩ thuật để khai phá dữ liệu LiDAR hiệu quả trong nghiên cứu các đối tượng trên bề mặt địa hình như EM, K-Means, kNN, MCC.*

Từ khóa*: data mining, LiDAR, phân loại, phân cụm, kNN.*

I. GIỚI THIỆU

Khai phá dữ liệu là quá trình khám phá những tri thức mới và tri thức có ích trong nguồn dữ liệu. Hiện nay, khai phá dữ liệu đang được ứng dụng trong rất nhiều lĩnh vực như thiên văn học, công nghệ thông tin, marketing, thể thao, giải trí,…. Khai phá dữ liệu là thuật ngữ mới xuất hiện từ những năm 2000 với sự bùng nổ của Internet và điện toán đám mây. Mục đích của khai phá dữ liệu là có thể trích xuất được những dữ liệu tiềm ẩn từ khối dữ liệu lớn được lưu trữ trong cơ sở dữ liệu, dựa trên những dữ liệu được khai phá này có thể cung cấp tri thức cho hệ hỗ trợ ra quyết định, đưa ra những thông tin dự báo hay có thể khái quát dữ liệu. Để thực hiện được những công việc này, khai phá dữ liệu có các phương pháp thực hiện quá trình khai phá như: phân loại, phân nhóm, tổng hợp, hồi quy, mô hình ràng buộc, dò tìm biến đổi và độ lệch. Với những bộ dữ liệu lớn, sử dụng các phương pháp rút trích thông tin thông thường như thống kê, máy học, hệ chuyên gia,... thường gặp khó khăn khi cơ sở dữ liệu trong thực tế thường thay đổi, nhiều dị thường và thường lớn hơn những bộ dữ liệu mẫu. Để có thể giải quyết được những khó khăn của bài toán cơ sở dữ liệu, khai phá dữ liệu dường như là công cụ hữu hiệu khi không phụ thuộc vào con người, bộ dữ liệu mẫu, hay các nhà thống kê [1].

LiDAR là công nghệ mới của ngành Trắc địa – Bản đồ, kể từ khi được áp dụng rộng rãi vào những năm 90 của thế kỉ trước, LiDAR đang ngày càng chứng tỏ được ưu thế của mình trong việc thu thập thông tin về những đối tượng trên bề mặt địa hình. Với khả năng thu nhận dữ liệu về một vùng rộng lớn, không bị hạn chế về thời tiết, đo được vào ban đêm, có khả năng đi xuyên qua nước và mặt đất, dữ liệu thu được từ hệ thống LiDAR là vô cùng lớn và có giá trị. Hệ thống LiDAR là một hệ thống tích hợp từ 3 thành phần chính: Hệ thống thiết bị Laser (Light amplification by stimulated emission of radiation), hệ thống định vị vệ tinh GNSS (Global Navigation Sattelite System) và hệ thống đạo hàng quán tính INS (Inertial Navigation System) [2]. Bản chất của công nghệ LiDAR là kỹ thuật đo dài laser, định vị không gian GPS/INS và sự nhận biết cường độ phản xạ ánh sáng. Xung của laser được phát hướng xuống mặt đất từ một độ cao nào đó. Sóng laser được phản hồi từ mặt đất hay từ các bề mặt đối tượng như là cây, đường hoặc nhà..., với mỗi xung sẽ đo được thời gian đi và về của tín hiệu, tính được khoảng cách từ nguồn phát laser tới đối tượng. 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. Trên cơ sở đó, sẽ tính được vị trí không gian của điểm phản xạ dưới dạng tọa độ (x, y, z) trong hệ tọa độ xác định nào đó. Một tín hiệu phát đi, sẽ có một hay nhiều tín hiệu phản xạ. Số lượng điểm phụ thuộc vào tính không gian của các đối tượng trên bề mặt đất [2]. Kết quả cuối cùng, sẽ có được đám mây điểm, số lượng điểm của đám mây điểm phụ thuộc vào sự phản xạ của tia laser khi gặp bề mặt các đối tượng trên bề mặt Trái đất, đám mây này thường rất lớn khoảng từ 4 - 5 triệu điểm. Với phương pháp xử lý số liệu thông thường được dùng trong Trắc địa truyền thống, để khai thác được những thông tin thu nhận được từ LiDAR vô cùng khó khăn, mất nhiều thời gian và công sức. Do đó, áp dụng những kĩ thuật khai phá dữ liệu trong bài toán xử lý dữ liệu LiDAR từ đó thu nhận được thông tin về các đối tượng trên bề mặt địa hình là hoàn toàn cần thiết và mang tính thực tiễn cao.

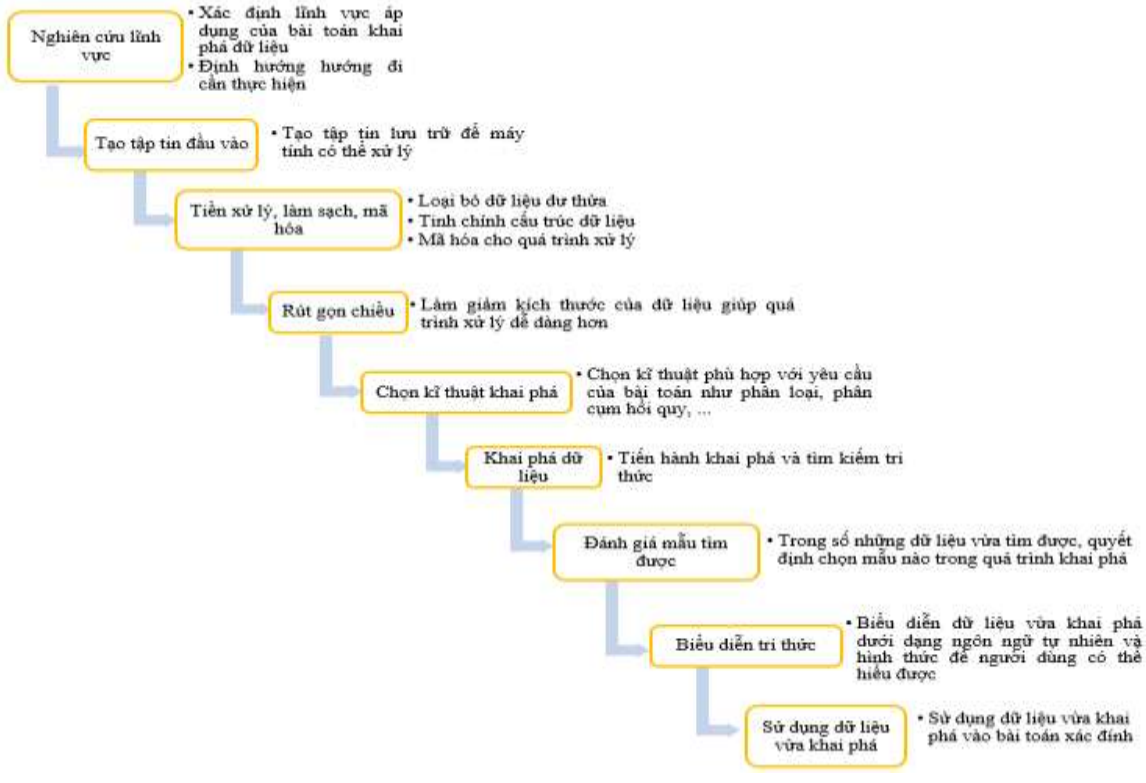
Ở Việt Nam, nghiên cứu các phương pháp khai phá dữ liệu LiDAR đã được các nhà khoa học công bố như nghiên cứu của TS. Trần Đình Luật, ThS Trần Đức Phú, GS.TSKH Lương Chính Kế đã nêu ra được các phương pháp sử dụng dữ liệu LiDAR để thành lập DTM, DSM, DEM mô tả về bề mặt địa hình của Trái đất trong các tài liệu [3] [4] [5].

Trên thế giới, khoa học công nghệ có những thành tựu phát triển vượt bậc, chính vì thế, công nghệ LiDAR được áp dụng từ khá sớm, có rất nhiều nghiên cứu của các nhà khoa học trên khắp thế giới về xử lý dữ liệu LiDAR như trong tài liệu tham khảo [6], [7], [8], [9], [10], … đã đưa ra các phương pháp phân loại, phân vùng dữ liệu LiDAR để từ đó nhận dạng các đối tượng bằng cách phân lớp cho đám mây điểm thu nhận được. Trong các nghiên cứu đó các tác giả đã trình bày các thuật toán phân cụm, phân loại đám mây điểm LiDAR truyền thống như K-Means, Maximum Likelihood, phân loại dựa trên điểm, … và một số thuật toán cải tiến hay dựa trên máy học như SVM (Support Vector Machine), EM (Expectation Maximization), MCC (Multiscale Curvature Classification), K-means theo thứ bậc…. Để thực hiện được quá trình khai phá dữ liệu LiDAR trong nghiên cứu các đối tượng trên bề mặt địa hình, nhóm tác giả đã nghiên cứu và đưa ra một số kĩ thuật cơ bản được trình bày trong phần [III] của bài báo. Trong phần [II] nhóm tác giả giới thiệu về tiến trình khai phá dữ liệu và một số kĩ thuật khai phá dữ liệu.

II. KỸ THUẬT KHAI PHÁ DỮ LIỆU

Tiến trình khai phá dữ liệu là sự kết hợp của các phương pháp, kỹ thuật và công cụ như thống kê, học máy, phát kiến khoa học, hệ chuyên gia để tìm ra một logic chính xác và phù hợp với cơ sở dữ liệu cần khai phá, từ mô hình này có thể ra những mâu thuẫn và quy tắc trong cơ sở dữ liệu. Tiến trình trích xuất cơ sở dữ liệu là một quá trình thao tác lặp đi lặp lại, không phải là một hệ thống phân tích tự động [1].

Tiến trình khai phá dữ liệu thường gồm các bước sau:



Hình 1. Tiến trình khai phá dữ liệu

Các giai đoạn của tiến trình khai phá dữ liệu là rất phức tạp, có thể lặp đi lặp lại rất nhiều lần, do đó, quá trình khai phá dữ liệu là phức tạp, phụ thuộc vào lĩnh vực áp dụng bài toán. Có nhiều kĩ thuật và thuật toán trong khai phá dữ liệu được nghiên cứu và thử nghiệm, trong đó các kĩ thuật khai phá dữ liệu chính như phân loại, phân nhóm, luật kết hợp, mạng nơ ron, k – Nearest Neighbor, hồi quy, phát hiện thay đổi và độ lệch. Mỗi kĩ thuật thích hợp với những bài toán cụ thể, có ưu và nhược điểm riêng trong khai phá dữ liệu. Trong bài báo này, để phù hợp với bài toán khai phá dữ liệu LiDAR nhóm tác giả tập trung trình bày kĩ thuật phân loại, phân nhóm và kNN.

- Phân loại (Classification): Là việc xác định một hàm ánh xạ từ một mẫu dữ liệu vào một trong số các lớp đã được biết trước đó. Mục tiêu của thuật toán phân lớp là tìm ra mối quan hệ nào đó giữa thuộc tính dự báo và thuộc tính phân lớp [1].

- Gom nhóm (Clustering): Là việc mô tả chung để tìm ra các tập hay các nhóm, loại mô tả dữ liệu. Các nhóm có thể tách nhau hoặc phân cấp hay gối lên nhau. Có nghĩa là dữ liệu có thể vừa thuộc nhóm này lại vừa thuộc nhóm khác [1].

- Hàng xóm gần nhất (k – Nearest Neighbor): 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. 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.

III. KHAI PHÁ DỮ LIỆU LIDAR TRONG NGHIÊN CỨU ĐỐI TƯỢNG TRÊN BỀ MẶT ĐỊA HÌNH

A. Đặc điểm của dữ liệu LiDAR

Dữ liệu đám mây điểm của LiDAR sẽ được tiền xử lý sau khi hệ thống thu nhận được hệ tọa độ x, y, z có độ chính xác cao của đối tượng bằng cách phân tích thời gian tia quét phản xạ, góc quét, vị trí thu nhận từ GPS và thông tin INS. Thuộc tính của dữ liệu LiDAR ghi nhận được cho mỗi xung bao gồm: cường độ, số lượng xung phản hồi, giá trị điểm phân loại, góc quét của đường bay chụp, thời gian định vị, góc quét và hướng quét.

Bảng 1. Thuộc tính của dữ liệu LiDAR *[11]*

STT	Tên thuộc tính	Mô tả
1	Cường độ	Độ đậm nhạt của xung dữ liệu LiDAR phản xạ ghi nhận được từ điểm LiDAR
2	Số lượng xung phản xạ	Tổng số lượng xung phản hồi
3	Điểm phân loại	Mọi điểm LiDAR đều được phân loại trong quá trình tiền xử lý để xác định được loại đối tượng phản xạ
4	Góc của đường quét	Các điểm sẽ được ký hiệu với giá trị 0 và 1. Những điểm theo đường góc quét sẽ được gán giá trị là 1, những điểm còn lại được gán giá trị là 0
5	RGB	Dữ liệu LiDAR có thể được gán với kênh phổ R, G, B. Giá trị này thường được thu nhận từ ảnh cùng thời gian với đo LiDAR
6	Thời gian định vị	Giờ được thu nhận từ hệ thống GPS được phát ra từ hệ thống không vận
7	Góc quét	Giá trị của góc quét thường từ - 90^0 đến + 90^0
8	Hướng quét	Là hướng gương chụp laser đang di chuyển tại thời điểm xung laser phát ra
9	x, y, z	Tọa độ và độ cao của điểm phản xạ

Dựa trên những thuộc tính này của dữ liệu LiDAR, tùy vào từng kĩ thuật và thuật toán sẽ lựa chọn sao cho phù hợp nhất với yêu cầu.

B. Nghiên cứu các đối tượng trên bề mặt địa hình bằng dữ liệu LiDAR sau phân loại

Bề mặt Trái đất có hình dạng gồ ghề, phức tạp, gồm các đại dương, lục địa và hải đảo. Địa hình của Trái đất tại mỗi lục địa và vị trí là khác nhau. Địa hình là phần mặt đất với các yếu tố trên bề mặt của nó như dáng đất, chất đất, thủy hệ, lớp thực vật, đường giao thông, điểm dân cư, các địa vật,…. Địa hình trên bề mặt Trái đất luôn thay đổi do có sự tác động của nội lực và ngoại lực, hai quá trình này ảnh hưởng đến nhau một cách nhất định, trong đó, nội lực đóng vai trò chủ yếu trong hình thành các yếu tố địa hình lớn, còn ngoại lực đóng vai trò trong hình thành các yếu tố địa hình nhỏ.

Các đối tượng trên bề mặt Trái đất gồm có:

- Địa vật: là những đối tượng của khu vực dễ dàng nhận biết trong khu vực đó.
- Thủy hệ: các đường bờ biển, bờ hồ, bờ sông lớn, hồ tự nhiên, hồ nhân tạo, …
- Điểm dân cư: được đặc trưng bởi kiểu cư trú, số người và ý nghĩa hành chính – chính trị của nó.
- Mạng lưới giao thông và đường dây liên lạc: đường quốc lộ, cấp tỉnh, cấp xã, huyện, …
- Dáng đất: được thể hiện thông qua các đường bình độ, mô hình độ cao, bề mặt, …
- Lớp phủ thực vật và đất: loại rừng, bụi cây, vườn cây, đồn điền, ruộng, đồng cỏ, ….
- Ranh giới hành chính: biên giới quốc gia, biên giới xã, huyện, tỉnh, ….

Với sự đa dạng về đối tượng và sự thay đổi liên tục của các nội dung địa hình, việc nghiên cứu là phức tạp và đòi hỏi sự đầu tư về thời gian cũng như công sức. Với công nghệ LiDAR, khả năng thu nhận dữ liệu về khu vực đo vẽ rộng lớn, độ chính xác cao, dễ triển khai và thực hiện đo vẽ, hoàn toàn thích hợp trong nghiên cứu các đối tượng trên bề mặt địa hình. Sự đa dạng của dữ liệu và khả năng ứng dụng trong nhiều bài toán cụ thể, LiDAR đang ngày càng được nghiên cứu và ứng dụng trong trắc địa – bản đồ. Trong nghiên cứu địa hình, ưu điểm nổi trội của dữ liệu LiDAR là có độ chính xác cao, thời gian thu nhận thông tin về địa hình được tiến hành rất nhanh với khoảng 4 – 8 điểm có tọa độ không gian ba chiều X, Y, Z trên một m^2. Mật độ điểm quét ở dưới dạng ô lưới với cạnh có thể điều chỉnh từ vài dm đến 10m [5].

1. Tạo DSM/DEM của bề mặt Trái đất

DEM/DSM là thuật ngữ chỉ các dữ liệu không gian liên tục ba chiều. DEM (Digital Elevation Model) – mô hình số độ cao là sự thể hiện độ cao của bề mặt Trái đất, độ cao của tầng đất của mặt nước, …. DSM là mô hình mô tả các đối tượng tự nhiên và nhân tạo trên bề mặt Trái đất, bao gồm các đối tượng địa vật trên địa hình như nhà cửa, cây cối, …. Để có thể thực hiện được điều này, từ dữ liệu đám mây điểm LiDAR, phải tiến hành phân loại dữ liệu thành hai lớp phản xạ đầu tiên (hay điểm không mặt đất – Non - Ground) và phản xạ cuối cùng (điểm mặt đất - Ground). Những

điểm thuộc vào lớp phản xạ cuối cùng sẽ được dùng để thành lập DEM, sử dụng cả hai lớp để tạo DSM. Thông thường DEM/DSM được tạo bằng cách phân loại, lọc các điểm độ cao trong đám mây điểm LiDAR để lựa chọn các điểm mặt đất và không mặt đất.

a) Phân loại đám mây điểm LiDAR tạo DEM

DEM là mô hình số độ cao thể hiện bằng số độ cao của bề mặt địa hình Trái đất, các dữ liệu liên quan đến độ cao này là liên tục thay đổi. DEM thông thường được sử dụng để có thể tìm hiểu về địa mạo và phân tích đứt gãy kiến tạo của khu vực nghiên cứu, có thể được sử dụng để tính toán độ dốc, sự lồi lõm của địa hình và các ứng dụng khác liên quan đến xử lý độ cao của đối tượng. Để tạo được DEM của bề mặt đất thực cần phải thực hiện công đoạn lọc và phân loại để có được độ cao thực tế của bề mặt địa hình. Để lọc được các điểm từ đám mây điểm LiDAR có thể sử dụng phương pháp chênh cao cực đại, độ võng cục bộ hay cực tiểu tuyệt đối.

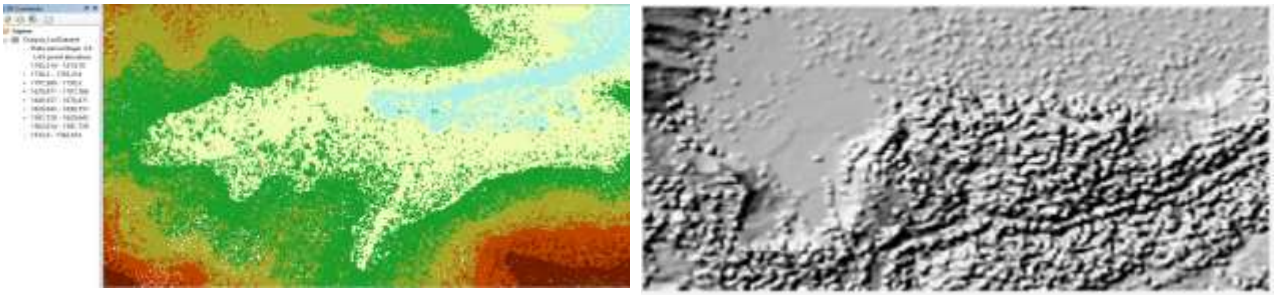
Theo [2] phương pháp phân loại đám mây điểm LiDAR thành hai lớp ground và non-ground có thể được tiến hành theo thuật toán chênh cao cực đại. Đây là thuật toán được sử dụng trong phân loại đám mây điểm LiDAR để thành lập DEM khá phổ biến trong các phần mềm thương mại như ENVI LiDAR, ArcGIS LDAR, …. Thuật toán được mô tả:

- Cho tập dữ liệu đầu vào P = {$p_1, \ldots, p_n$}

- Điểm p_i được lọc từ tập hợp A là tập hợp điểm địa hình nếu ∄ p_j nào khác, sao cho ($h_{pi} - h_{pj}$) < Δh_{max} trong phạm vi khoảng cách giữa hai điểm d(p_i , p_j) với góc quét Asc

- Tiến hành phân loại điểm từ p_{min}, sau đó tiến hành sắp xếp tăng dần các điểm, điểm địa hình sẽ được lọc theo các ngưỡng k_1, k_2 trong đó $k_1 < k_2$. Ta có chuẩn lọc như sau:

(a) ($h_i - h_{min}$) < k_1/cosb => là điểm địa hình
(b) ($h_i - h_{min}$) > k_2/cosb => là điểm địa vật
(c) k_1/cosb < ($h_i - h_{min}$) < k_2/cosb => là điểm không được lọc (không được phân loại)



(a) Các điểm sau phân loại

(b) DEM được thành lập sau phân loại điểm LiDAR

Hình 2. Thành lập DEM với lớp điểm mặt đất

Sau quá trình lọc các điểm địa hình và địa vật theo độ cao với ngưỡng cho trước, ta phân chia các điểm độ cao theo từng nhóm (trong hình (a)) và tiến hành thành lập DEM của khu vực với điểm độ cao vừa lọc (trong hình (b)).

b) Phân loại đám mây điểm LiDAR tạo DSM

DSM là mô hình mô tả cho độ cao của bề mặt vật lý Trái đất, nó được thành lập từ tín hiệu phản xạ đầu tiên của đám xung phản xạ LiDAR. DSM ngoài bao gồm các yếu địa hình còn bao gồm các yếu tố địa vật như cây cối, đường dây điện và toàn nhà. DSM khác với DEM ở chỗ DEM cung cấp thông tin về độ cao trên bề mặt thật của Trái đất không bao gồm các địa vật, trong khi đó DSM bao gồm các đối tượng tự nhiên. DSM sau khi được thành lập có thể được sử dụng để mô hình hóa bề mặt cảnh quan, mô hình hóa đô thị và các ứng dụng trực quan, kiểm tra lớp phủ và một số ứng dụng thông thường khác như đo khoảng cách, ….

Từ đám mây điểm LiDAR để tạo được DSM ta tiến hành chọn những điểm có độ cao cao nhất từ các điểm, để thực hiện điều này cần tiến hành phân loại đám mây điểm thành hai lớp tia phản hồi đầu tiên (First Pulse - FP) và tia phản hồi cuối cùng (Last Pulse - LP), các điểm còn lại không thuộc hai lớp này đều bị loại bỏ.

Ta có thể sử dụng phương pháp phân nhóm các đối tượng sử dụng thuật toán phân cụm K-means. K-means là thuật toán phân cụm dựa trên khoảng cách phổ biến, với nghiên cứu trong tài liệu [12], tác giả Kun Zhang và nnk đã sử dụng thuật toán K-means để phân loại đám mây điểm. Với thuật toán này, tác giả nhận thấy rằng thời gian để thực hiện thuật toán nhanh hơn, nhưng độ chính xác vẫn còn phụ thuộc nhiều vào số cụm khởi tạo. Để tiến hành phân chia đám mây điểm thành hai lớp FP và LP, tiến hành như sau:

Hình 3. Mô tả thuật toán K-means *[13]*

Với bộ dữ liệu đám mây điểm với 3 thuộc tính là (x, y, z), trong đó thuộc tính được sử dụng để phân cụm là z (độ cao của điểm), kết quả chạy với phần mềm thống kê của IBM - SPSS, lựa chọn k = 2.

Initial Cluster Centers

	Cluster	
	1	2
Chieu cao Z	112,58	193,77

(a) Tâm khởi tạo của cụm

Final Cluster Centers

	Cluster	
	1	2
Chieu cao Z	138,16	168,41

(b) Tâm của cụm trong lần lặp cuối

Number of Cases in each Cluster

Cluster	1	139,000
	2	61,000
Valid		200,000
Missing		,000

(c) Sự phân bố của các cụm sau khi quá trình lặp kết thúc

Hình 4. Phân cụm đám mây điểm LiDAR với k = 2

Hai cụm được phân chia sau quá trình phân cụm với cụm 1 có 139 điểm, cụm 2 có 61 điểm trong tổng số 200 điểm, với tâm của cụm 1 là điểm có độ cao 138,16; cụm 2 là 168,41. Số điểm được gán vào cụm số 1 là 139 trong tổng số 200 điểm, cụm số 2 là 61 điểm. Giá trị Missing (điểm không được gán vào cụm nào) là 0.

2. Phân loại lớp phủ mặt đất sử dụng dữ liệu LiDAR

Lớp phủ mặt đất là lớp phủ quan sát được khi nhìn từ mặt đất hoặc thông qua một số phương pháp đo đạc như viễn thám, quang học như thực vật (tự nhiên hoặc nhân tạo), các công trình xây dựng trên đất (nhà, công trình giao thông, …), mặt nước, …. Trên thực tế, mỗi khu vực trên Trái đất đều có loại hình lớp phủ mặt đất đặc trưng và mỗi đối tượng chịu sự tác động của tự nhiên và con người là khác nhau, chính sự tác động này làm cho lớp phủ mặt đất luôn biến đổi. Để nghiên cứu và tìm hiểu về khu vực đo vẽ, tìm hiểu được loại lớp phủ và đặc trưng của nó giúp cho công việc thuận lợi hơn. Để thu thập được thông tin về lớp phủ mặt đất tại khu vực đo vẽ phương pháp sử dụng tư liệu viễn thám là phương pháp hiện đại, giúp trích xuất thông tin về lớp phủ nhanh chóng, hiệu quả và công nghệ LiDAR là công nghệ viễn thám được sử dụng để phân loại lớp phủ mặt đất. Với những khu vực có kết cấu phức tạp như đô thị, thường phân loại lớp phủ mặt đất khá phức tạp do có nhiều đối tượng, nhiều thuộc tính lựa chọn.

a) Phân loại lớp phủ mặt đất trong khu vực đô thị

Lớp phủ mặt đất ở khu vực đô thị luôn biến đổi không ngừng do sự tác động của đô thị hóa. Để phân loại lớp phủ khu vực đô thị ta có thể sử dụng DSM/DEM được thành lập từ LIDAR. Do DSM/DEM được thành lập từ công nghệ LiDAR nhanh, có độ chính xác cao, tự động và có khả năng trích xuất được những thông tin về đối tượng trên bề mặt ở khu vực đô thị, nên đây là phương pháp hoàn toàn khả thi.

Phương pháp có thể được tiến hành như sau [14]:

- Tiến hành phân loại không giám sát với mô hình DSM và sử dụng bộ lọc trung bình để phát hiện ra luật phân bố của bề mặt

- Nhận dạng mẫu

- Tính toán luật phân bố chuẩn σ để phát hiện ra quy luật của bề mặt gốc

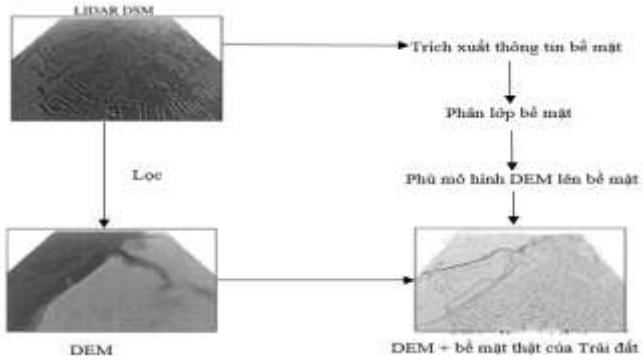
- Tạo một bề mặt tham chiếu gần với bề mặt thật của khu vực khảo sát nhất

- Tạo lớp đặc trưng bề mặt gốc từ bề mặt tham chiếu

- Tạo DEM từ bề mặt tham chiếu và DSM

- Sử dụng thuật toán để phân loại các đối tượng ở khu vực đô thị

- Từ quá trình phân loại, trên khu vực đo vẽ ta có thể tiến hành nghiên cứu và chọn các đối tượng trong lớp theo yêu cầu.



Hình 5. Phân loại lớp phủ mặt đất dựa trên LiDAR DSM (Nguồn: *[14]*)

Với lớp phủ mặt đất khu vực đô thị, đây là khu vực có độ phức tạp, luôn biến động và chịu sự tác động của những hoạt động của con người. Thông tin được trích xuất từ quá trình phân loại có thể được sử dụng cho quy hoạch đô thị trung hạn và dài hạn, kiểm soát sự phát triển của cơ sở hạ tầng và trang thiết bị xã hội, đồng thời nó cũng hỗ trợ cho quá trình bảo vệ môi trường sống của thành phố. Phương pháp phân loại lớp phủ mặt đất ở khu đô thị được thực hiện với cây phân loại, giúp quá trình trích lọc và trích xuất thông tin về các đối tượng trên bề mặt đô thị được thực hiện [15]:

- Cho bộ dữ liệu đầu vào là đám mây điểm LiDAR và ảnh viễn thám về khu vực đo vẽ

- Tiến hành tiền xử lý dữ liệu để loại bỏ những dữ liệu dư thừa và lựa chọn bộ dữ liệu trong quá trình huấn luyện

- Lựa chọn các đặc trưng trong quá trình phân đoạn dữ liệu sử dụng thuật toán di truyền theo công thức:

$$P_{obt} = arg_P \min([F(S,P)]) \quad (2)$$

Trong đó, S là đoạn tham chiếu, F là hàm so sánh độ lệch của các đoạn

- Sử dụng kĩ thuật phân loại theo cây quyết định để phân loại các lớp trong mỗi đoạn.

Với cây quyết định, mỗi nút thể hiện cho thuộc tính phân loại, tại mỗi nút thuộc tính được chọn là thuộc tính tốt nhất để chia dữ liệu thành các lớp riêng biệt.

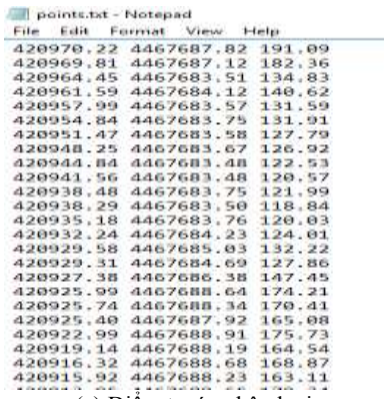
- Cuối cùng, các lớp sẽ được gán nhãn với thuộc tính ở nút tương ứng mà nó thuộc về.

b) Phân loại lớp phủ thực vật mặt đất

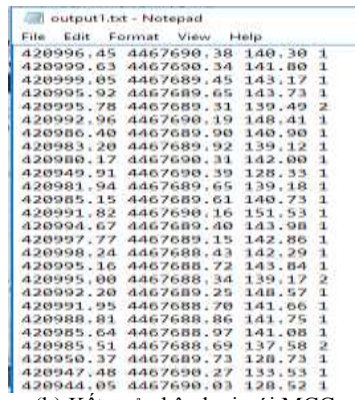
Lớp phủ thực vật mặt đất là toàn bộ thảm thực vật xuất hiện trên mặt đất bao gồm thực vật mọc tự nhiên và thực vật được trồng do con người [16]. Từ những thông tin trích xuất được từ lớp phủ thực vật ta có thể biết được tình trạng tài nguyên thực vật của vùng, thể hiện của đặc điểm tự nhiên của vùng, phản ánh đa dạng sinh học, sự phân bố của các đối tượng tự nhiên trên bề mặt nghiên cứu. Thành phần của lớp phủ thực vật như cỏ, thực vật thấp, cây bụi, cây lá kim, cây lâu năm, rừng trồng, lúa, dân cư,…. Với sự thay đổi thường xuyên của bề mặt lớp phủ và sự phân bố ở những địa hình phức tạp, việc áp dụng công nghệ LiDAR trong phân loại lớp phủ thực vật phần nào giúp giải quyết khó khăn cho bài toán này.

Bài toán phân loại lớp phủ thực vật sử dụng thuật toán MCC (Multiscale Curvature Classification) là thuật toán phân loại dựa trên số liệu (x, y, z) của đám mây điểm LiDAR. Thuật toán MCC là thuật toán được Evans và Hudak đề xuất vào năm 2007, chủ yếu được sử dụng để phân loại đám mây điểm LiDAR trong môi trường có nhiều thực vật. Thuật toán có ý tưởng phân loại dựa trên ngưỡng độ cong và một phép lặp TPS (Thin Plate Spline) để gán một điểm vào lớp mặt đất hay không mặt đất. Thuật toán được mô tả như sau:

Input: đám mây điểm LiDAR
Output: điểm sau phân loại với hai lớp mặt đất và không mặt đất
Initial: U_0 đám mây điểm LiDAR
$U = \{P_1, P_2, \ldots, P_n\}$ – chứa các điểm chưa được phân loại
$P_j(x_j, y_j, z_j)$ – điểm LiDAR rời rạc
int l[1 … 3] – miền tỉ lệ
t_l – độ cong cho miền tỉ lệ l – khởi tạo bởi người dùng
λ – độ phân giải cho l
Procedure:
Phân loại những điểm có độ cao trong U_0 và gán chúng vào lớp không mặt đất
For l = 1 to 3
Repeat
Lặp tìm S = TPS(U, λ, f)
Lặp tìm S' = 3x3
For each $P_j \in U$
If $z_j > S'(x_j, y_j) + t_{li}$ then
Phân P_j vào lớp không mặt đất - Loại P_j khỏi U
Untill điểm không thuộc U < 10% tổng số điểm của nó
Phân lớp các điểm còn lại của U vào lớp mặt đất
End



(a) Điểm trước phân loại



(b) Kết quả phân loại với MCC

Hình 6. Phân loại đám mây điểm với MCC

Đám mây điểm sau khi được phân loại với MCC sẽ được gán với mã 1 nếu là điểm thuộc lớp không mặt đất, mã 2 với điểm thuộc lớp mặt đất. Từ mã lớp này có thể lựa chọn được các điểm thực vật để từ đó tiến hành phân chia thành các cụm thực vật tương ứng.

Hoặc có thể tiến hành phân loại theo thuật toán kNN. Thuật toán kNN là thuật toán phân loại có giám sát, phân loại dựa trên khoảng cách của điểm cần phân loại đến tập dữ liệu mẫu. Độ chính xác của thuật toán phụ thuộc hoàn toàn vào khoảng cách. Đây là thuật toán tương đối đơn giản, dễ cài đặt.

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

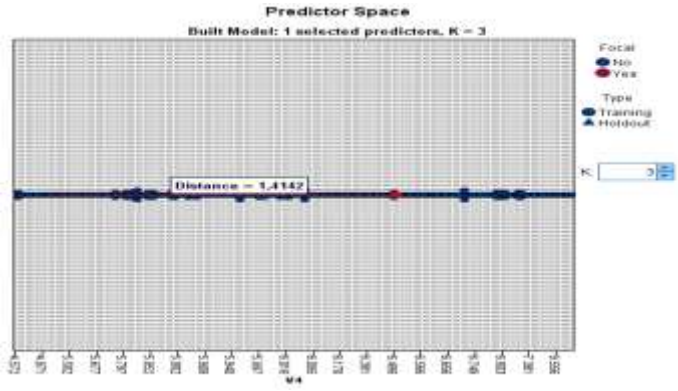
Tìm k điểm tương ứng với k giá trị

If $k_i > k_j \ \forall \ i \neq j$ then

Gán điểm p_i vào lớp i

End

Với đám mây điểm cho trước với 353 điểm, phân loại kNN với phần mềm SPSS cho kết quả mô hình dự đoán phân loại như sau:



(a) Chọn điểm

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

(b) Kết quả điểm gần nhất và khoảng cách gần nhất

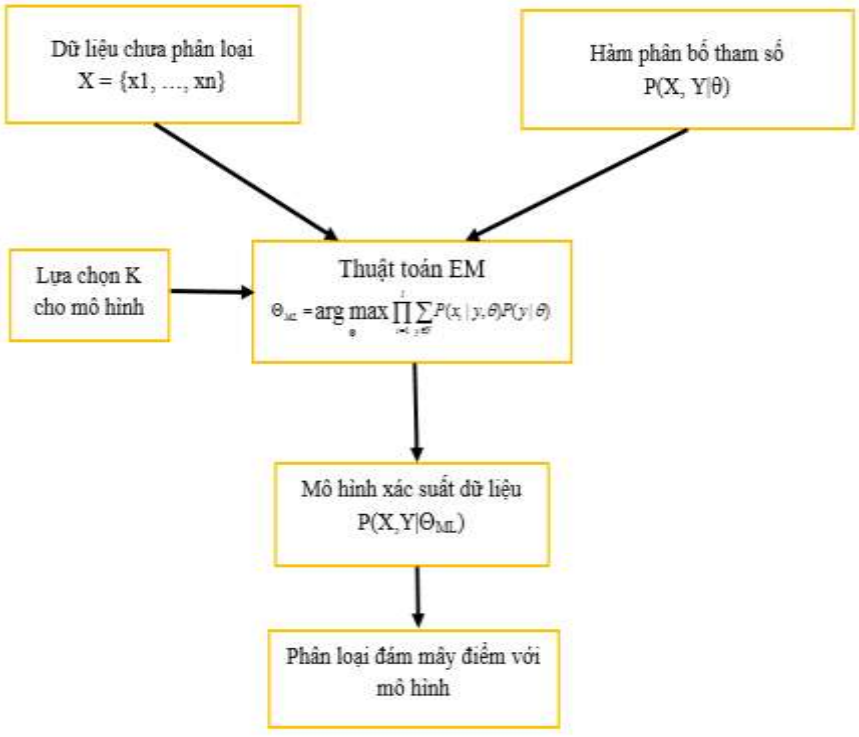
Hình 7. Mô hình dự báo phân loại theo kNN với SPSS

Kết quả phân loại theo kNN được thể hiện trong hình 10. Lựa chọn giá trị K = 3, điểm cần dự báo phân loại là 230, dựa trên khoảng cách chương trình tìm ra được 3 điểm hàng xóm gần nhất với nó dựa trên khoảng cách gần nhất được thể hiện trong hình (b).

3. Phân cụm dữ liệu bề mặt Trái đất

Với khu vực cần nghiên cứu, sự sắp xếp phức tạp của các đối tượng gây khó khăn cho các nhà nghiên cứu trong quá trình chọn lựa và trích xuất đối tượng. Phân tích cấu trúc của bề mặt địa hình của Trái đất được thực hiện nhằm làm giảm bớt các khó khăn trong quá trình nghiên cứu. Do những ưu điểm của mình, dữ liệu LiDAR được sử dụng như là dữ liệu chính của quá trình này. Phân cụm bề mặt Trái đất là quá trình chia bề mặt Trái đất thành các lớp và nhóm các điểm trong không gian.

Thuật toán EM (Expectation Maximization) là thuật toán được sử dụng để ước tính khả năng ước lượng tối đa cho mô hình tham số khi dữ liệu chưa đầy đủ, thiếu dữ liệu, hoặc có các biến tiềm ẩn không quan sát được. Thuật toán EM có thể tìm mô hình tham số ngay cả trong bộ dữ liệu không đầy đủ. Thuật toán chọn giá trị ngẫu nhiên cho dữ liệu bị thiếu và sử dụng những dự đoán để tính toán được bộ thứ hai của dữ liệu. Các giá trị mới sẽ được sử dụng để tạo ra một dự đoán tốt hơn cho tập dữ liệu đầu và quá trình này sẽ lặp đi lặp lại cho đến khi thuật toán hội tụ [17]. Để thực hiện phân cụm các đối tượng trên bề mặt Trái đất theo sơ đồ sau:



Hình 8. Mô tả thuật toán EM trong phân loại đám mây điểm LiDAR

(a) Dữ liệu gốc

(b) Dữ liệu sau khi tiến hành phân cụm với màu đỏ là khu vực đô thị, màu xám là đường giao thông, đất nông thôn là màu vàng, thực vật màu xanh

Hình 9. Phân cụm các đối tượng trên bề mặt đo vẽ (Nguồn *[18]*)

IV. KẾT LUẬN

Dữ liệu LiDAR kể từ khi được áp dụng trong ngành Khoa học Trái đất và Trắc địa - Bản đồ đã thể hiện được ưu điểm và vai trò của mình. Với khả năng thu thập dữ liệu về khu vực rộng lớn, đo được vào ban đêm và không bị ảnh hưởng bởi thời tiết, có khả năng đi xuyên đối tượng để thu thập thông tin. Chính vì lý do đó, hiện nay LiDAR đang được sử dụng và coi như là dữ liệu chính trong các nghiên cứu về các đối tượng trên bề mặt. Trong phần [III] của bài báo, nhóm tác giả đã chỉ ra được khả năng của dữ liệu LiDAR sau khi sử dụng các kĩ thuật của khai phá dữ liệu để phân loại và phân cụm trong việc trích xuất thông tin sử dụng cho nghiên cứu, quy hoạch, dự đoán, … bề mặt địa hình. Dữ liệu LiDAR sau khai phá là bộ dữ liệu lớn, rất có ý nghĩa nghĩa khi cần nghiên cứu bề mặt không gian của đô thị, sinh khối của thảm thực vật, trích xuất thông tin về đối tượng và sử dụng cho những bài toán quy hoạch trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Khoa Công nghệ thông tin, Bài giảng Khai phá dữ liệu dành cho Cao học, Hải Phòng: Trường Đại học Hàng hải, 2011.

[2] T. Đ. Trí, Công nghệ LiDAR, Hà Nội: Trường Đại học Mỏ - Địa chất, 2013.

[3] T. Đ. Phú, "Ứng dụng công nghệ LiDAR trong mô hình hóa lũ," *Tạp chí Khoa học công nghệ và hàng hải,* vol. 23, pp. 54-58, 2010.

[4] Trần Đình Luật, Nguyễn Thị Kim Dung, Lưu Thị Thu Thủy, Trần Hồng Hạnh, "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,* vol. 1, pp. 24-28, 2015.

[5] T. L. C. Kế, "Thành lập DEM/DTM DSM bằng công nghệ LiDAR," 2005.

[6] Bao Yunfei, Li Guoping, Cao Chunxiang, Li Xiaowen, Zhang Hao, "Classification of LiDAR point cloud and generation of DTM form LiDAR heigth and intensity data in forest area," *The International Archives of the Photogrammetry, RS and Spatial Information Sciences,* vol. XXXVII, no. B3b, pp. 314-318, 2008.

[7] A. Brzank, C. Heipke, "Classificatin of LiDAR data into water and land points in coastal areas," 2007.

[8] D. Gajski, I. Fiedler, A. Krtalic, "Classification and filtering of airborne topographic LiDAR data," 2004.

[9] S. Filin, "Surface clustering from airborne laser scanning data," 2001.

[10] Madhurina Bandyopadhyay, Jan A. N van Aardt, Kerry Cause - Nicholson, "Classification and extraction of trees and buildings from urban scenes using discrete return LiDAR and aerial color imagery," *Laser Radar Technology and Applications,* vol. 8731, 2013.

[11] ESRI, "http://desktop.arcgis.com/en/arcmap/10.3/manage-data/las-dataset/what-is-lidar-data-.htm," 2016. [Online].

[12] Kun Zhang, Weihong Bi, Xiaoming Zhang, Xinghu Fu, Kunpeng Zhu, Li Zhu, "A new kmeans clustering algorithm for point cloud," *International Journal of Hybrid Information Technology,* vol. 8, no. 9, pp. 157-170, 2015.

[13] Nguyễn Thị Hữu Phương, Đặng Văn Đức, Nguyễn Trường Xuân, "Sử dụng thuật toán K-means trong phân loại đám mây điểm LiDAR," *Tạp chí Khoa học và Công nghệ,* vol. 5, no. 5B, pp. 1-5, 2017.

[14] G. Priestnall, J. Jaafar, A. Duncan, "Extracting urban features from LiDAR digital surfaces model," *Computers, Environments and Urban Systems,* vol. 24, pp. 65-78, 2000.

[15] F. Leonardi, C. M. Almeida, nnk, "Genetic algorithms and data mining applied to optical orbital and LiDAR data for object-based classification of urban land cover," in *4th GEOBIA*, Rio de Janeiro, Brazil, 2012.

[16] H. V. Thuân, "Đại học Thái Nguyên," Đại học Thái Nguyên, 4 2012. [Online]. Available: http://qlkh.tnu.edu.vn/theme/details/1337/ket-hop-thong-tin-tu-anh-ve-tinh-da-pho-da-thoi-gian-bang-phuong-phap-thong-ke-da-bien-de-nang-ca. [Accessed 31 3 2017].

[17] Alex Berson, nnk, "An overview of data mining technique," *Building DM Applications for CRM*, 2005.

[18] Jorge Garcia Gutierrez, Francisco Martinez Alvarez, Jose C.Riquelme, "Using remote data mining on LiDAR and Imagery Fusion data to develop Land Cover maps," *IEA/AIE*, vol. I, pp. 378-387, 2010.

LIDAR DATA DEVELOPMENT IN THE STUDY OF SUBJECTS ON THE SURFACE

Nguyen Thi Huu Phuong, Dang Van Duc, Nguyen Truong Xuan

ABSTRACT: *LiDAR, since its inception in the 1990s, has been applied to many sectors and many areas of social life. With the ability to receive input data to the large scan area, measured at night and not affected by weather, the data collected from LiDAR is of high precision. Compared with the conventional geodetic techniques, LiDAR has outstanding advantages in the study of subjects on the surface. However, the data obtained from the LiDAR system is relatively large, with a range of 4000 to 5000 points per square kilometer including coordinates, sweep time, elevation, and so on. In order to use LiDAR data effectively in the study of objects on the surface, it is necessary to have data mining techniques in order to find useful information. This paper focuses on presenting techniques for mining LiDAR data effectively in the study of objects on the surface.*