TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO SINH HOẠT HỌC THUẬT

"Công nghệ Thông tin Địa học và Trí tuệ Nhân Tạo"

(Geospatial Artificial Intelligence)

Người báo cáo: Trần Trường Giang

Bộ môn: Tin học – Trắc địa

HÀ NỘI, 12 -2021

MỤC LỤC

1. Giới thiệu về Công nghệ Thông tin Địa học

Công nghệ thông tin địa học được hiểu là sự kết hợp tri thức và công nghệ của các lĩnh vực Viễn Thám (Remote Sensing - RS), Hệ Thống Thông Tin Địa Lý (Geographic Information Systems - GIS) và Hệ Thống Vệ Tinh Điều Hướng Toàn Cầu (Global Navigation Satellite System - GNSS) được sử dụng trong rất nhiều ứng dụng trong kĩ thuật và dân sinh.

Công nghệ Thông tin Địa học cung cấp các công cụ hiệu quả để thu thập thông tin cần thiết cho bảo vệ môi trường, giám sát tài nguyên trái đất và giải quyết các vấn đề liên quan đến quản lý và phát triển bền vững. Các công cụ này được tạo ra chuyên biệt hóa cho các vấn đề xử lý cấu trúc và đặc điểm của thông tin không gian (spatial information), thu thập thông tin, phân loại và chất lượng, lưu trữ, xử lý, mô tả và phổ biến, bao gồm cả cơ sở hạ tầng cần thiết để đảm bảo việc sử dụng thông tin này một cách tối ưu.

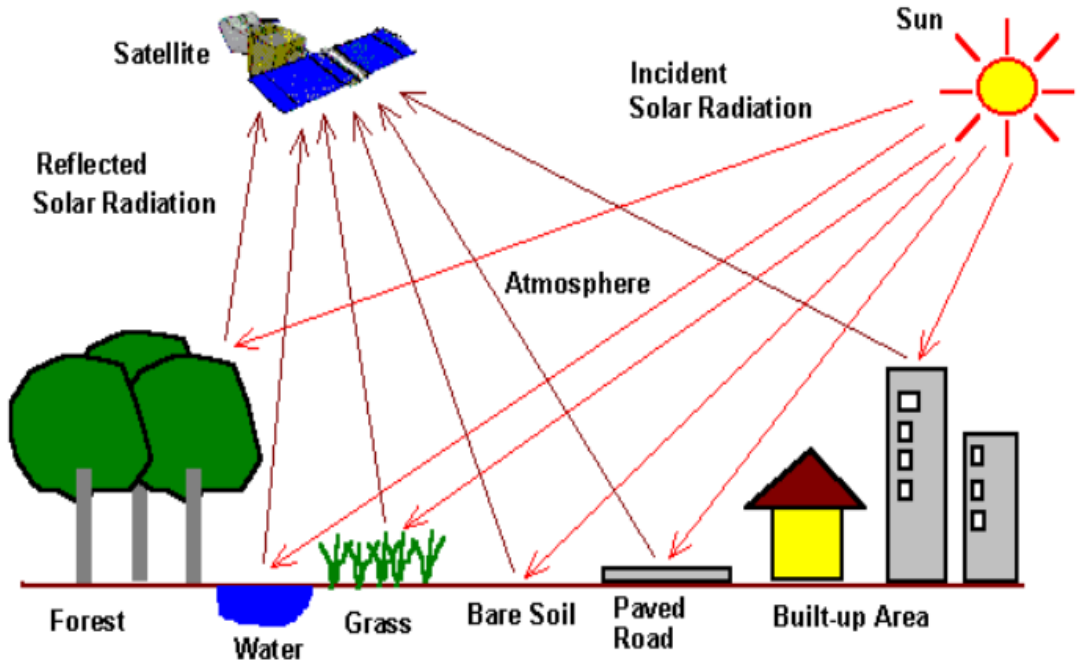
Hình 1: Công nghệ Thông tin Địa học

1.1 Công nghệ Viễn thám (RS)

Viễn thám có thể được chia thành hai loại phương pháp: Viễn thám thụ động và Viễn thám chủ động. Cảm biến thụ động thu thập bức xạ được phát ra hoặc phản xạ bởi đối tượng hoặc các khu vực xung quanh. Ánh sáng mặt trời phản xạ là nguồn bức xạ phổ biến nhất được đo bằng cảm biến thụ động. Ví dụ về cảm biến từ xa thụ động bao gồm chụp ảnh phim, hồng ngoại, thiết bị tích hợp điện tích và máy đo bức xạ. Mặt khác, bộ thu tích cực phát ra năng lượng để quét các đối tượng và khu vực mà sau đó cảm biến phát hiện và đo bức xạ bị phản xạ hoặc tán xạ ngược từ mục tiêu. RADAR và LiDAR là những ví dụ về viễn thám chủ động, trong đó thời gian trễ giữa phát xạ và quay lại được đo, thiết lập vị trí, tốc độ và hướng của một đối tượng.

Viễn thám cho phép thu thập dữ liệu của các khu vực nguy hiểm hoặc không thể tiếp cận. Các ứng dụng viễn thám bao gồm giám sát nạn phá rừng ở các khu vực như Lưu vực sông Amazon, các đặc điểm băng giá ở các vùng Bắc Cực và Nam Cực, và đo độ sâu của độ sâu ven biển và đại dương. Bộ sưu tập quân sự trong Chiến tranh Lạnh đã sử dụng việc thu thập dữ liệu dự phòng về các khu vực biên giới nguy hiểm.

Viễn thám cũng thay thế việc thu thập dữ liệu tốn kém và chậm chạp trên mặt đất, đảm bảo trong quá trình này các khu vực hoặc đối tượng không bị xáo trộn.



Hình 2: Công nghệ Viễn thám

Các nền tảng quỹ đạo thu thập và truyền dữ liệu từ các phần khác nhau của phổ điện từ, kết hợp với cảm biến và phân tích trên không hoặc trên mặt đất quy mô lớn hơn, cung cấp cho các nhà nghiên cứu đủ thông tin để theo dõi các xu hướng như El Niño và các hiện tượng tự nhiên dài hạn và ngắn hạn khác. Các mục đích sử dụng khác bao gồm các lĩnh vực khác nhau của khoa học trái đất như quản lý tài nguyên thiên nhiên, các lĩnh vực nông nghiệp như sử dụng và bảo tồn đất, giám sát khí nhà kính, phát hiện và giám sát sự cố tràn dầu, và an ninh quốc gia và thu gom trên cao, trên mặt đất và thu gom tại các khu vực biên giới.

Để tạo bản đồ dựa trên cảm biến, hầu hết các hệ thống viễn thám đều mong đợi ngoại suy dữ liệu cảm biến liên quan đến điểm tham chiếu bao gồm khoảng cách giữa các điểm đã biết trên mặt đất. Điều này phụ thuộc vào loại cảm biến được sử dụng. Ví dụ, trong các bức ảnh thông thường, khoảng cách là chính xác ở trung tâm của hình ảnh, với sự biến dạng của các phép đo càng tăng càng xa trung tâm. Một yếu tố khác là trục cuốn mà phim được ép lên có thể gây ra lỗi nghiêm trọng khi ảnh được sử dụng để đo khoảng cách mặt đất. Bước giải quyết vấn đề này được gọi là tham chiếu địa lý và liên quan đến việc kết hợp các điểm trong hình ảnh với sự hỗ trợ của máy tính (thường là 30 điểm trở lên trên mỗi hình ảnh), được ngoại suy với việc sử dụng một điểm chuẩn đã thiết lập, "làm cong" hình ảnh để tạo ra độ chính xác dữ liệu không gian. Vào đầu những năm 1990, hầu hết các hình ảnh vệ tinh được bán tham chiếu địa lý đầy đủ.

Ngoài ra, hình ảnh có thể cần được hiệu chỉnh bằng bức xạ và khí quyển. **Hiệu chỉnh bức xạ** (Radiometric Correction): Cho phép tránh các lỗi và biến dạng đo bức xạ. Sự chiếu sáng của các vật thể trên bề mặt Trái đất là không đồng đều vì các đặc tính khác nhau của phù điêu. Yếu tố này được tính đến trong phương pháp hiệu chỉnh biến dạng bức xạ. Hiệu chỉnh đo bức xạ cung cấp tỷ lệ cho các giá trị pixel, ví dụ: thang đo đơn sắc từ 0 đến 255 sẽ được chuyển đổi thành các giá trị bức xạ thực tế. **Hiệu chỉnh địa hình** (Topographic Correction): Ở những vùng núi hiểm trở, do địa hình, độ chiếu

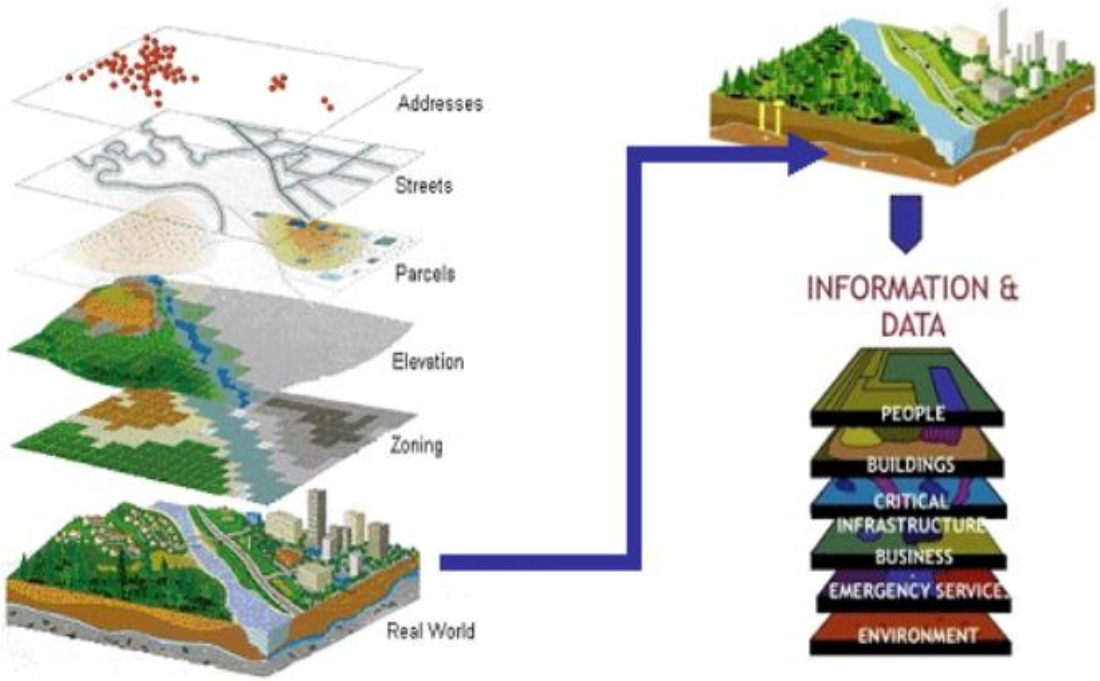
sáng hiệu quả của các pixel thay đổi đáng kể. Trong ảnh viễn thám, điểm ảnh trên dốc có bóng râm nhận được ánh sáng yếu và có giá trị bức xạ thấp, ngược lại, điểm ảnh trên dốc có nắng nhận được ánh sáng mạnh và có giá trị bức xạ cao. Đối với cùng một đối tượng, giá trị bức xạ pixel trên mái dốc râm sẽ khác với giá trị trên mái dốc có nắng. Ngoài ra, các đối tượng khác nhau có thể có các giá trị bức xạ tương tự. Những điều không rõ ràng này đã ảnh hưởng nghiêm trọng đến độ chính xác của việc trích xuất thông tin ảnh viễn thám ở các khu vực miền núi. Nó trở thành trở ngại chính cho việc ứng dụng ảnh viễn thám. Mục đích của việc hiệu chỉnh địa hình là loại bỏ hiệu ứng này, khôi phục hệ số phản xạ hoặc độ tỏa sáng thực của các đối tượng trong điều kiện nằm ngang. Nó là tiền đề của ứng dụng viễn thám định lượng. **Hiệu chỉnh khí quyển** (Atmospheric correction): Loại bỏ sương mù trong khí quyển bằng cách thay đổi tỷ lệ từng dải tần sao cho giá trị nhỏ nhất của nó (thường được nhận ra ở các vùng nước) tương ứng với giá trị pixel bằng 0. Việc số hóa dữ liệu cũng giúp bạn có thể thao tác dữ liệu bằng cách thay đổi các giá trị thang xám .

Nói chung, viễn thám hoạt động theo nguyên tắc của bài toán ngược: trong khi đối tượng hoặc hiện tượng quan tâm (trạng thái) có thể không được đo trực tiếp, nhưng tồn tại một số biến số khác có thể được phát hiện và đo lường (quan sát) có thể liên quan cho đối tượng quan tâm thông qua một phép tính. Phép loại suy phổ biến được đưa ra để mô tả điều này là cố gắng xác định loại động vật từ dấu chân của nó. Ví dụ, trong khi không thể đo trực tiếp nhiệt độ ở tầng trên của bầu khí quyển, có thể đo phổ phát xạ từ một loại hóa chất đã biết (chẳng hạn như carbon dioxide) trong vùng đó. Sau đó, tần số của phát xạ có thể liên quan thông qua nhiệt động lực học với nhiệt độ trong vùng đó.

1.2 Hệ thống Thông tin Địa lý (GIS)

Hệ thống thông tin địa lý (GIS) là một loại cơ sở dữ liệu chứa dữ liệu địa lý (nghĩa là mô tả các hiện tượng có liên quan đến vị trí), kết hợp với các công cụ phần mềm để quản lý, phân tích và trực quan hóa các dữ liệu đó. Theo nghĩa rộng hơn, người ta có thể coi một hệ thống như vậy cũng bao gồm người dùng và nhân viên hỗ trợ, các thủ tục và quy trình làm việc, khối kiến thức về các khái niệm và phương pháp liên quan, và các tổ chức thể chế. Hệ thống thông tin địa lý được sử dụng trong nhiều công nghệ, quy trình, kỹ thuật và phương pháp. Chúng gắn liền với các hoạt động khác nhau và nhiều ứng dụng liên quan đến: kỹ thuật, lập kế hoạch, quản lý, vận tải / hậu cần, bảo hiểm, viễn thông và kinh doanh. Vì lý do này, GIS và các ứng dụng thông minh vị trí là nền tảng của các dịch vụ hỗ trợ vị trí, dựa trên phân tích địa lý và trực quan hóa. GIS cung cấp khả năng liên hệ các thông tin không liên quan trước đây, thông qua việc sử dụng vị trí làm "biến chỉ số chính". Các vị trí và phạm vi được tìm thấy trong không thời gian của Trái đất có thể được ghi lại thông qua ngày và thời gian xuất hiện, cùng với các tọa độ x, y và z; đại diện, kinh độ (x), vĩ độ (y) và độ cao (z). Tất cả các tham chiếu dựa trên Trái đất, không gian-thời gian, vị trí và phạm vi phải liên quan với nhau, và cuối cùng, với một vị trí hoặc phạm vi thực. Đặc tính quan trọng này của GIS đã bắt đầu mở ra những con đường mới cho việc tìm hiểu và nghiên cứu khoa học.

Cốt lõi của bất kỳ hệ thống GIS nào là cơ sở dữ liệu chứa các đại diện của các hiện tượng địa lý, mô hình hóa hình học (vị trí và hình dạng) và các thuộc tính hoặc đặc tính của chúng. Cơ sở dữ liệu GIS có thể được lưu trữ ở nhiều dạng khác nhau, chẳng hạn như một tập hợp các tệp dữ liệu riêng biệt hoặc một cơ sở dữ liệu quan hệ hỗ trợ không gian duy nhất. Việc thu thập và quản lý những dữ liệu này thường bao gồm phần lớn thời gian và nguồn lực tài chính của một dự án, nhiều hơn nhiều so với các khía cạnh khác như phân tích và lập bản đồ.



Hình 3: Hệ thống Thông tin Địa lý

Khía cạnh dữ liệu địa lý: GIS sử dụng vị trí không gian-thời gian (không-thời gian) làm biến chỉ số chính cho tất cả các thông tin khác. Cũng giống như cơ sở dữ liệu quan hệ có chứa văn bản hoặc số có thể liên hệ nhiều bảng khác nhau bằng cách sử dụng các biến chỉ số khóa chung, GIS có thể liên hệ thông tin không liên quan bằng cách sử dụng vị trí làm biến chỉ số chính. Điều quan trọng là vị trí và / hoặc phạm vi trong không-thời gian. Liên quan đến thông tin không gian chính xác, nhiều loại dữ liệu trong quá khứ hoặc tương lai được dự báo trong thế giới thực và dự kiến có thể được phân tích, giải thích và biểu diễn. Đặc điểm chính này của GIS đã bắt đầu mở ra những con đường nghiên cứu khoa học mới về các hành vi và mẫu thông tin trong thế giới thực mà trước đây không có mối tương quan một cách hệ thống.

Khía cạnh mô hình hóa dữ liệu: Dữ liệu GIS đại diện cho các hiện tượng tồn tại trong thế giới thực, chẳng hạn như đường xá, sử dụng đất, độ cao, cây cối, đường thủy và các trạng thái. Các loại hiện tượng phổ biến nhất được biểu diễn trong dữ liệu có thể được chia thành hai khái niệm: các đối tượng rời rạc (ví dụ: một ngôi nhà, một con đường) và các trường liên tục (ví dụ: lượng mưa hoặc mật độ dân số). Các loại hiện tượng địa lý khác, chẳng hạn như các sự kiện (ví dụ: Chiến tranh thế giới thứ hai), các quá trình (ví dụ: ngoại ô hóa) và khối lượng (ví dụ: loại đất) được thể hiện ít phổ biến hơn hoặc gián tiếp, hoặc được mô hình hóa trong các thủ tục phân tích hơn là dữ liệu. Theo truyền thống, có hai phương pháp rộng được sử dụng để lưu trữ dữ liệu trong GIS cho cả hai loại tham chiếu ánh xạ trừu tượng: hình ảnh raster và vectơ. Điểm, đường và đa giác đại diện cho dữ liệu vectơ của các tham chiếu thuộc tính vị trí được ánh xạ. Một phương pháp lưu trữ dữ liệu kết hợp mới là xác định các đám mây điểm, kết hợp các điểm ba chiều với thông tin RGB tại mỗi điểm, trả về "hình ảnh màu 3D". Các bản đồ chuyên đề GIS sau đó ngày càng trở nên mô tả trực quan thực tế hơn về những gì họ đặt ra để hiển thị hoặc xác định.

Khía cạnh thu thập dữ liệu: Thu thập dữ liệu GIS bao gồm một số phương pháp thu thập dữ liệu không gian vào cơ sở dữ liệu GIS, có thể được nhóm thành ba loại: thu

thập dữ liệu chính, các hiện tượng đo trực tiếp tại hiện trường (ví dụ, viễn thám, hệ thống định vị toàn cầu); thu thập dữ liệu thứ cấp, trích xuất thông tin từ các nguồn hiện có không ở dạng GIS, chẳng hạn như bản đồ giấy, thông qua số hóa; và truyền dữ liệu, sao chép dữ liệu GIS hiện có từ các nguồn bên ngoài như các cơ quan chính phủ và các công ty tư nhân. Tất cả các phương pháp này có thể tiêu tốn đáng kể thời gian, tài chính và các nguồn lực khác.

Khía cạnh hệ tham chiếu không gian: Trái đất có thể được biểu thị bằng nhiều mô hình khác nhau, mỗi mô hình có thể cung cấp một tập hợp tọa độ khác nhau (ví dụ: vĩ độ, kinh độ, độ cao) cho bất kỳ điểm nhất định nào trên bề mặt Trái đất. Mô hình đơn giản nhất là giả định trái đất là một hình cầu hoàn hảo. Khi nhiều phép đo về trái đất được tích lũy, các mô hình của trái đất ngày càng tinh vi và chính xác hơn. Trên thực tế, có những mô hình được gọi là datums áp dụng cho các khu vực khác nhau của trái đất để tăng độ chính xác, như Datum Bắc Mỹ năm 1983 cho các phép đo của Hoa Kỳ và Hệ thống trắc địa thế giới cho các phép đo trên toàn thế giới. Kinh độ và vĩ độ trên bản đồ được tạo dựa trên dữ liệu địa phương có thể không giống với kinh độ và vĩ độ thu được từ máy thu GPS. Việc chuyển đổi tọa độ từ mức dữ liệu này sang mức dữ liệu khác yêu cầu phép chuyển đổi mức dữ liệu chẳng hạn như phép biến đổi Helmert, mặc dù trong một số trường hợp nhất định, một phép dịch đơn giản có thể là đủ. Trong phần mềm GIS phổ biến, dữ liệu chiếu theo vĩ độ / kinh độ thường được biểu diễn dưới dạng hệ tọa độ Địa lý. Ví dụ: dữ liệu ở vĩ độ / kinh độ nếu dữ liệu là 'Dữ liệu thống kê Bắc Mỹ năm 1983' được ký hiệu là 'GCS Bắc Mỹ năm 1983'.

Khía cạnh chất lượng dữ liệu: Mặc dù không có mô hình kỹ thuật số nào có thể thể hiện hoàn hảo thế giới thực, nhưng điều quan trọng là dữ liệu GIS phải có chất lượng cao. Để phù hợp với nguyên tắc tương đồng, dữ liệu phải đủ gần với thực tế để kết quả của các thủ tục GIS tương ứng chính xác với kết quả của các quá trình trong thế giới thực. Điều này có nghĩa là không có tiêu chuẩn duy nhất cho chất lượng dữ liệu, bởi vì mức độ cần thiết của chất lượng phụ thuộc vào quy mô và mục đích của nhiệm vụ mà nó được sử dụng. Một số yếu tố của chất lượng dữ liệu rất quan trọng đối với dữ liệu GIS: **Độ chính xác** (accuracy) - Mức độ giống nhau giữa một phép đo được đại diện và giá trị thực tế; ngược lại, sai số là số lượng khác biệt giữa chúng. Trong dữ liệu GIS, người ta quan tâm đến độ chính xác trong các biểu diễn về vị trí (độ chính xác của vị trí), thuộc tính (độ chính xác của thuộc tính) và thời gian. Ví dụ, Điều tra dân số Hoa Kỳ năm 2020 nói rằng dân số của Houston vào ngày 1 tháng 4 năm 2020 là 2.304.580; nếu nó thực sự là 2.310.674, đây sẽ là một lỗi và do đó thiếu độ chính xác của thuộc tính. **Độ tin cậy** (precision): Mức độ tinh chỉnh trong một giá trị được đại diện. Trong thuộc tính định lượng, đây là số chữ số có nghĩa trong giá trị đo được. Giá trị không chính xác là mơ hồ hoặc không rõ ràng, bao gồm một loạt các giá trị có thể có. Ví dụ: nếu người ta nói rằng dân số của Houston vào ngày 1 tháng 4 năm 2020 là "khoảng 2,3 triệu", tuyên bố này sẽ không chính xác, nhưng có thể chính xác vì giá trị đúng (và nhiều giá trị không chính xác) được bao gồm. Cũng như độ chính xác, các đại diện về vị trí, tài sản và thời gian đều có thể chính xác hơn hoặc ít hơn. Độ phân giải là một biểu thức thường được sử dụng của độ chính xác vị trí, đặc biệt là trong các tập dữ liệu raster. **Độ bất định** (uncertainty): Sự thừa nhận chung về sự hiện diện của lỗi và sự không chính xác trong dữ liệu địa lý. Đó là mức độ nghi ngờ chung, do rất khó để biết chính xác có bao nhiêu lỗi trong tập dữ liệu, mặc dù một số hình thức ước lượng có thể được thử (khoảng tin cậy là một ước lượng về độ không đảm bảo). Điều này đôi khi được sử dụng như một thuật ngữ chung cho tất cả hoặc hầu hết các khía cạnh của chất lượng dữ liệu. **Độ mờ** (fuzziness): Mức độ mà một khía cạnh (vị trí, đặc tính hoặc thời gian) của một hiện tượng vốn dĩ không chính xác, thay vì không chính xác ở một giá

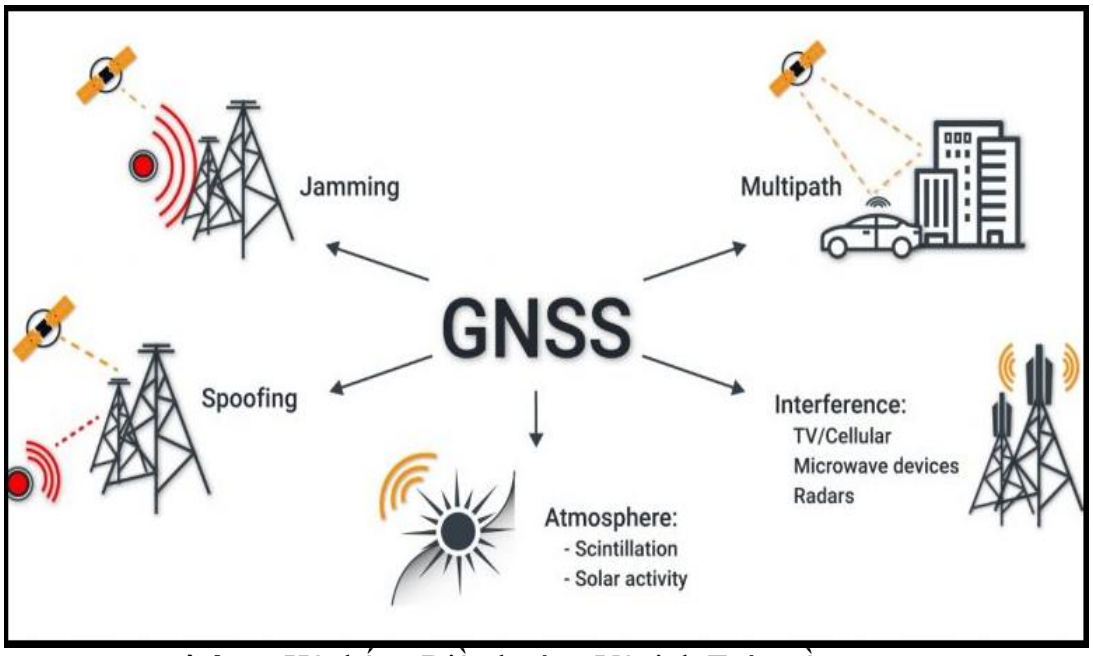
trị đo lường. Ví dụ, phạm vi không gian của đô thị Houston khu vực này không rõ ràng, vì có những nơi ở ngoại ô thành phố ít được kết nối với thành phố trung tâm (được đo bằng các hoạt động như đi lại) hơn những nơi gần hơn. Các công cụ toán học như lý thuyết tập mờ thường được sử dụng để quản lý sự mơ hồ trong dữ liệu địa lý. Độ hoàn chỉnh (completeness): Mức độ mà một tập dữ liệu đại diện cho tất cả các đối tượng địa lý thực tế mà nó muốn đưa vào. Ví dụ: nếu một lớp "đường ở Houston" thiếu một số đường phố thực tế, thì nó chưa hoàn chỉnh. **Độ cập nhật** (currency): Thời điểm gần đây nhất mà tại đó tập dữ liệu tuyên bố là đại diện chính xác của thực tế. Đây là mối quan tâm đối với phần lớn các ứng dụng GIS, những ứng dụng cố gắng đại diện cho thế giới "hiện tại", trong trường hợp đó, dữ liệu cũ có chất lượng thấp hơn. **Độ nhất quán** (consistency): Mức độ mà các biểu diễn của nhiều hiện tượng trong tập dữ liệu tương ứng chính xác với nhau. Tính nhất quán trong các mối quan hệ tôpô giữa các đối tượng không gian là một khía cạnh đặc biệt quan trọng của tính nhất quán. Ví dụ: nếu tất cả các đường trong mạng lưới đường phố vô tình bị dịch chuyển 10 mét về phía Đông, chúng sẽ không chính xác nhưng vẫn nhất quán, vì chúng vẫn kết nối đúng cách tại mỗi giao lộ và các công cụ phân tích mạng như đường đi ngắn nhất vẫn cho kết quả chính xác.

Phân tích không gian trong GIS là một lĩnh vực thay đổi nhanh chóng và các gói GIS ngày càng bao gồm các công cụ phân tích dưới dạng cơ sở tích hợp tiêu chuẩn, dưới dạng bộ công cụ tùy chọn, như phần bổ trợ hoặc 'nhà phân tích'. Trong nhiều trường hợp, chúng được cung cấp bởi các nhà cung cấp phần mềm gốc (nhà cung cấp thương mại hoặc các nhóm phát triển phi thương mại hợp tác), trong khi trong các trường hợp khác, các cơ sở đã được phát triển và được cung cấp bởi các bên thứ ba. Hơn nữa, nhiều sản phẩm cung cấp bộ công cụ phát triển phần mềm (SDK), ngôn ngữ lập trình và hỗ trợ ngôn ngữ, phương tiện viết kịch bản và / hoặc giao diện đặc biệt để phát triển các công cụ hoặc biến thể phân tích của riêng mình. Tính khả dụng ngày càng tăng đã tạo ra một khía cạnh mới cho trí thông minh kinh doanh được gọi là "trí thông minh không gian", khi được phân phối công khai qua mạng nội bộ, dân chủ hóa quyền truy cập vào dữ liệu mạng xã hội và địa lý. Trí thông minh không gian địa lý, dựa trên phân tích không gian GIS, cũng trở thành một yếu tố quan trọng để bảo mật. GIS nói chung có thể được mô tả là chuyển đổi sang biểu diễn vector hoặc sang bất kỳ quá trình số hóa nào khác. Quá trình xử lý không gian (geoprocessing) là một loạt các tác vụ GIS được sử dụng để xử lý dữ liệu không gian. Một hoạt động xử lý địa lý điển hình lấy một tập dữ liệu đầu vào, thực hiện một hoạt động trên tập dữ liệu đó và trả về kết quả của hoạt động dưới dạng một tập dữ liệu đầu ra. Các hoạt động xử lý địa lý phổ biến bao gồm lớp phủ đối tượng địa lý, lựa chọn và phân tích đối tượng địa lý, xử lý cấu trúc liên kết, xử lý raster và chuyển đổi dữ liệu. Xử lý địa lý cho phép xác định, quản lý và phân tích thông tin được sử dụng để hình thành các quyết định.

1.3 Hệ thống Vệ tinh Điều hướng Toàn cầu (GNSS)

Hệ thống Vệ tinh Điều hướng Toàn cầu (global navigation satellite system - GNSS) là một hệ thống sử dụng vệ tinh để cung cấp định vị không gian địa lý tự trị. Nó cho phép các máy thu điện tử nhỏ xác định vị trí của chúng (kinh độ, vĩ độ và độ cao / độ cao) với độ chính xác cao (trong vòng vài cm đến mét) bằng cách sử dụng tín hiệu thời gian được truyền dọc theo đường ngắm bằng sóng vô tuyến từ vệ tinh. Hệ thống có thể được sử dụng để cung cấp vị trí, điều hướng hoặc theo dõi vị trí của một thứ gì đó được gắn với bộ thu (theo dõi vệ tinh). Các tín hiệu cũng cho phép bộ thu điện tử tính toán giờ địa phương hiện tại với độ chính xác cao, cho phép đồng bộ hóa thời gian. Những công dụng này được gọi chung là Định vị, Điều hướng và Định thời (PNT). Hệ

thống GNSS hoạt động độc lập với bất kỳ hệ thống thu nhận điện thoại hoặc internet, mặc dù những công nghệ này có thể nâng cao tính hữu ích của thông tin định vị được tạo ra.



Hình 4: Hệ thống Điều hướng Vệ tinh Toàn cầu

Hệ thống GNSS cung cấp độ chính xác nâng cao và toàn vẹn cho quá trình giám sát có thể sử dụng cho các ứng dụng dân sự được phân loại như sau: **GNSS-1**: là hệ thống thế hệ đầu tiên và là sự kết hợp của các hệ thống định vị vệ tinh hiện có (GPS và GLONASS), với các trạm mặt đất được phân bố nhiều nơi trên trái đất như các trạm WAAS (Wide Area Augmentation System) ở Châu Mỹ, các trạm EGNOS (European Geostationary Navigation Overlay Service) ở Châu Âu, các trạm MSAS (Multi-Functional Satellite Augmentation System) ở Nhật Bản. **GNSS-2**: là thế hệ thứ hai của hệ thống cung cấp độc lập hệ thống định vị vệ tinh dân sự đầy đủ, ví dụ như hệ thống định vị Galileo của Châu Âu. Các hệ thống này sẽ cung cấp độ chính xác và giám sát toàn vẹn cần thiết cho các ứng dụng dân dụng, kể cả máy bay. Ban đầu, hệ thống này chỉ bao gồm các bộ tần số trên Dải tần L (L1 cho GPS, E1 cho Galileo, G1 cho GLONASS). Trong những năm gần đây, các hệ thống GNSS đã bắt đầu kích hoạt các bộ tần số L-Band thấp hơn (L2 và L5 cho GPS, E5a và E5b cho Galileo, G3 cho GLONASS) cho mục đích dân dụng. Chúng có độ chính xác tổng hợp cao hơn và ít vấn đề hơn với phản xạ tín hiệu. Tính đến cuối năm 2018, một số thiết bị GNSS cấp dành cho người tiêu dùng đang được bán tận dụng cả hai và thường được gọi là thiết bị "GNSS băng tần kép" hoặc "GPS băng tần kép". Do nhiều hệ thống GNSS toàn cầu (và các hệ thống nâng cấp) sử dụng các tần số và tín hiệu tương tự xung quanh L1, nhiều máy thu "Multi-GNSS" có khả năng sử dụng nhiều hệ thống đã được sản xuất. Trong khi một số hệ thống cố gắng tương thích tốt nhất với GPS bằng cách cung cấp cùng một đồng hồ, những hệ thống khác thì không.

Động lực ban đầu của định vị vệ tinh là dành cho các ứng dụng quân sự. Điều hướng qua vệ tinh cho phép sự chính xác trong việc đưa vũ khí tới mục tiêu, làm tăng đáng kể khả năng sát thương của chúng đồng thời giảm thương vong do vũ khí điều hướng sai. (Xem Bom có hướng dẫn). Định vị vệ tinh cũng cho phép các lực lượng được định hướng và xác định vị trí của họ dễ dàng hơn, giảm sương mù chiến tranh.

Giờ đây, một hệ thống vệ tinh định vị toàn cầu, chẳng hạn như Galileo, được sử dụng để xác định vị trí của người dùng và vị trí của người hoặc vật thể khác tại bất kỳ thời điểm nào. Phạm vi ứng dụng của định vị vệ tinh trong tương lai là rất lớn, bao gồm cả khu vực công và tư nhân trên nhiều phân khúc thị trường như khoa học, giao thông, nông nghiệp. Khả năng cung cấp tín hiệu định vị vệ tinh cũng là khả năng phủ nhận tính khả dụng của chúng. Người vận hành hệ thống định vị vệ tinh có khả năng làm suy giảm hoặc loại bỏ các dịch vụ định vị vệ tinh trên bất kỳ lãnh thổ nào mà họ mong muốn.

GPS (Global Positioning System) - Hệ thống Định vị Toàn cầu (GPS) của Hoa Kỳ bao gồm tối đa 32 vệ tinh quỹ đạo Trái đất tầm trung trong sáu mặt phẳng quỹ đạo khác nhau. Số lượng chính xác của các vệ tinh thay đổi khi các vệ tinh cũ hơn bị loại bỏ và thay thế. Hoạt động từ năm 1978 và có mặt trên toàn cầu từ năm 1994, GPS là hệ thống định vị vệ tinh được sử dụng nhiều nhất trên thế giới.

GLONASS (GLObal NAvigation Satellite System) - Hệ thống định vị dựa trên vệ tinh không gian của Liên Xô trước đây và hiện nay là của Nga cung cấp dịch vụ định vị vệ tinh vô tuyến dân sự và cũng được sử dụng bởi Lực lượng Phòng vệ Hàng không Vũ trụ Nga. GLONASS có phạm vi phủ sóng toàn cầu từ năm 1995 và với 24 vệ tinh.

Galileo - Vào tháng 3 năm 2002, Liên minh Châu Âu và Cơ quan Vũ trụ Châu Âu đã đồng ý đưa ra giải pháp thay thế GPS cho riêng mình, được gọi là hệ thống định vị Galileo. Galileo bắt đầu hoạt động vào ngày 15 tháng 12 năm 2016 (Khả năng hoạt động sớm toàn cầu, EOC). Với chi phí ước tính khoảng 10 tỷ €, hệ thống gồm 30 vệ tinh MEO ban đầu được lên kế hoạch hoạt động vào năm 2010. Năm đầu tiên bắt đầu hoạt động là năm 2014. Vệ tinh thử nghiệm đầu tiên được phóng vào ngày 28 tháng 12 năm 2005. Galileo dự kiến sẽ tương thích với hệ thống GPS hiện đại. Các máy thu sẽ có thể kết hợp các tín hiệu từ cả vệ tinh Galileo và GPS để tăng độ chính xác lên rất nhiều. Chòm sao Galileo đầy đủ bao gồm 24 vệ tinh đang hoạt động, vệ tinh cuối cùng được phóng vào tháng 12 năm 2021. Điều chế chính được sử dụng trong tín hiệu Dịch vụ mở Galileo là điều chế sóng mang bù lệch nhị phân tổng hợp (CBOC).

Bắc Đẩu – Bắc đẩu bắt đầu với tên gọi Bắc đẩu -1 (Beidou-1) hiện đã ngừng hoạt động, một mạng cục bộ châu Á - Thái Bình Dương trên quỹ đạo địa tĩnh. Thế hệ thứ hai của hệ thống có tên Bắc đẩu – 2 (BeiDou-2) bắt đầu hoạt động tại Trung Quốc vào tháng 12 năm 2011. Hệ thống Bắc đẩu – 3 (BeiDou-3) được đề xuất bao gồm 30 vệ tinh MEO và 5 vệ tinh địa tĩnh (IGSO). Một phiên bản khu vực gồm 16 vệ tinh (bao gồm khu vực Châu Á và Thái Bình Dương) đã được hoàn thành vào tháng 12 năm 2012. Dịch vụ toàn cầu được hoàn thành vào tháng 12 năm 2018. Vào ngày 23 tháng 6 năm 2020, việc triển khai chòm sao BDS-3 được hoàn tất sau khi vệ tinh cuối cùng được phóng thành công tại Trung tâm phóng vệ tinh Tây Xương.

Các ứng dụng của GNSS có thể kể ra như: **Điều hướng** (navigation): Ô tô có thể được trang bị bộ thu GNSS tại nhà máy hoặc dưới dạng thiết bị hậu mãi. Các đơn vị thường hiển thị bản đồ di chuyển và thông tin về vị trí, tốc độ, hướng, các đường phố lân cận và các điểm ưa thích. Hệ thống định vị hàng không thường có màn hình hiển thị bản đồ di chuyển và thường được kết nối với hệ thống lái tự động để điều hướng trên đường bay. Máy thu GNSS gắn trong buồng lái và buồng lái bằng kính đang xuất hiện trên các máy bay hàng không nói chung ở mọi kích cỡ, sử dụng các công nghệ như WAAS hoặc LAAS để tăng độ chính xác. Nhiều chiếc được chứng nhận về điều hướng các quy tắc bay bằng thiết bị, và một số cũng có thể được sử dụng cho các hoạt động tiếp cận và hạ cánh cuối cùng như trong hệ thống tiếp cận và hạ cánh chính xác chung. Các phi công lái tàu lượn sử dụng Máy ghi chuyến bay GNSS để ghi dữ liệu GNSS xác minh việc họ đến điểm rẽ trong các cuộc thi lái tàu lượn và để biết thông tin hỗ trợ việc

ra quyết định trên đường bay xuyên quốc gia. Thuyền và tàu có thể sử dụng GNSS để điều hướng tất cả các hồ, biển và đại dương trên thế giới. Các đơn vị GNSS hàng hải bao gồm các chức năng hữu ích trên mặt nước, chẳng hạn như chức năng "người trên tàu" (MOB) cho phép đánh dấu ngay lập tức vị trí nơi một người rơi xuống tàu, giúp đơn giản hóa nỗ lực cứu hộ. GNSS có thể được kết nối với thiết bị tự lái của tàu và Máy vẽ biểu đồ bằng giao diện NMEA 0183. GNSS cũng có thể cải thiện tính bảo mật của lưu lượng vận chuyển bằng cách kích hoạt AIS. **Đo đạc và thành lập bản đồ** (surveying and mapping): Hầu hết các máy thu GNSS cấp ánh xạ chỉ sử dụng dữ liệu sóng mang từ tần số L1, nhưng có bộ dao động tinh thể chính xác giúp giảm các lỗi liên quan đến rung nhịp đồng hồ máy thu. Điều này cho phép sai số định vị theo thứ tự từ một mét trở xuống trong thời gian thực, với tín hiệu GNSS vi sai nhận được bằng cách sử dụng một máy thu vô tuyến riêng biệt. Bằng cách lưu trữ các phép đo pha sóng mang và xử lý sau vi phân dữ liệu, các lỗi định vị theo thứ tự 10 cm có thể xảy ra với các máy thu này. Máy thu GNSS cấp độ khảo sát có thể được sử dụng để định vị các điểm đánh dấu khảo sát, các tòa nhà và xây dựng đường. Các thiết bị này sử dụng tín hiệu từ cả tần số GPS L1 và L2. Mặc dù dữ liệu mã L2 đã được mã hóa, sóng mang của tín hiệu vẫn cho phép sửa một số lỗi tầng điện ly. Các máy thu GPS tần số kép này thường có giá từ 10.000 đô la Mỹ trở lên, nhưng có thể có sai số định vị theo thứ tự từ một cm trở xuống khi được sử dụng ở chế độ GPS vi phân pha sóng mang.

2. Giới thiệu Trí tuệ Nhân tạo

Trí tuệ nhân tạo (AI) là trí thông minh được thể hiện bởi máy móc, trái ngược với trí thông minh tự nhiên được hiển thị bởi động vật bao gồm cả con người. Nghiên cứu AI được định nghĩa là lĩnh vực nghiên cứu các tác nhân thông minh, đề cập đến bất kỳ hệ thống nào nhận thức được môi trường của nó và thực hiện các hành động nhằm tối ưu hóa cơ hội đạt được mục tiêu của nó. Thuật ngữ "trí tuệ nhân tạo" trước đây đã được sử dụng để mô tả các máy móc bắt chước và hiển thị các kỹ năng nhận thức của "con người" gắn liền với trí óc con người, chẳng hạn như "học tập" (learning) và "giải quyết vấn đề" (problem solving). Định nghĩa này đã bị bác bỏ bởi các nhà nghiên cứu AI lớn, những người hiện đang mô tả AI về tính hợp lý (rationality) và hành động hợp lý (acting rationally), điều này không giới hạn cách trí thông minh có thể được khớp nối.

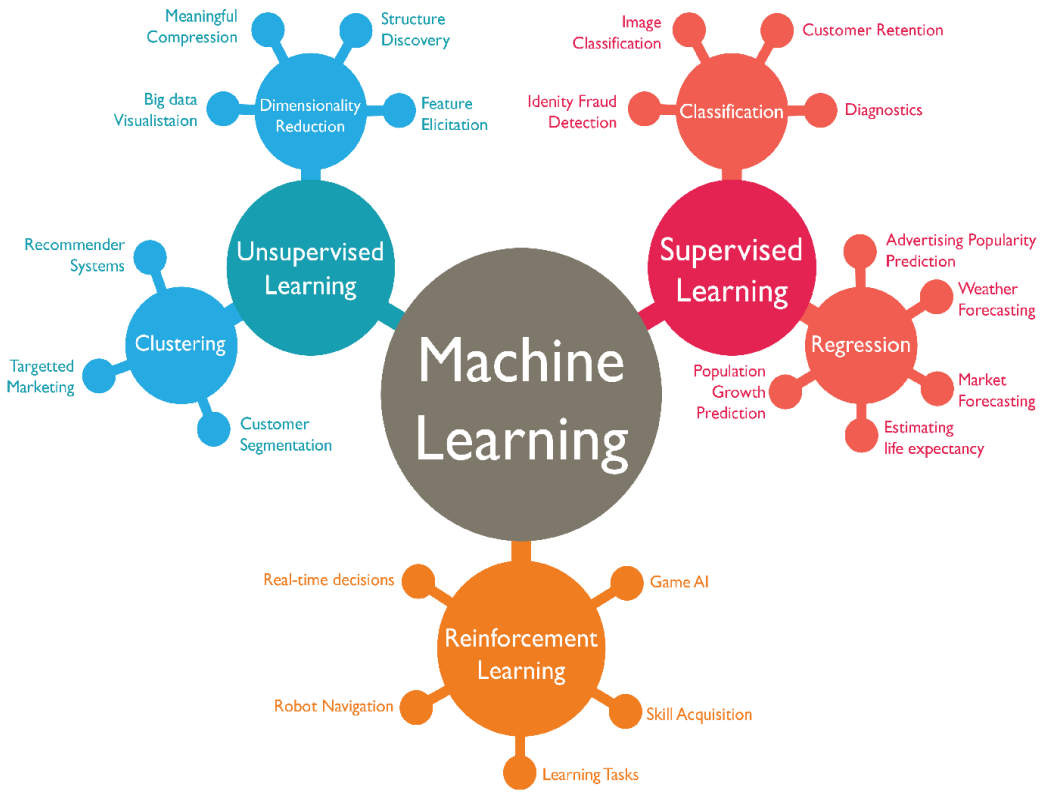
Các ứng dụng AI bao gồm công cụ tìm kiếm web nâng cao (ví dụ: Google), hệ thống đề xuất (được sử dụng bởi YouTube, Amazon và Netflix), hiểu giọng nói của con người (chẳng hạn như Siri và Alexa), ô tô tự lái (ví dụ: Tesla), ra quyết định tự động và cạnh tranh ở cấp độ cao nhất trong các hệ thống trò chơi chiến lược (chẳng hạn như cờ vua và cờ vây). Khi máy móc ngày càng trở nên có năng lực, các nhiệm vụ được coi là đòi hỏi "trí thông minh" thường bị loại bỏ khỏi định nghĩa về AI, một hiện tượng được biết đến như hiệu ứng AI. Ví dụ, nhận dạng ký tự quang học thường bị loại trừ khỏi những thứ được coi là AI, đã trở thành một công nghệ thông thường.

Trí tuệ nhân tạo được thành lập như một bộ môn học thuật vào năm 1956, và trong những năm kể từ đó đã trải qua một số làn sóng lạc quan, tiếp theo là sự thất vọng và mất nguồn tài chính (được gọi là "mùa đông AI"), tiếp theo là các cách tiếp cận mới, thành công và nguồn tài trợ mới. Nghiên cứu AI đã thử và loại bỏ nhiều cách tiếp cận khác nhau kể từ khi thành lập, bao gồm mô phỏng bộ não, mô hình hóa giải quyết vấn đề của con người, logic chính thức, cơ sở dữ

liệu lớn về kiến thức và bắt chước hành vi của động vật. Trong những thập kỷ đầu tiên của thế kỷ 21, các thuật toán Máy Học Tập (machine learning) dựa trên toán thống kê đã thống trị lĩnh trí tuệ nhân tạo, và kỹ thuật Máy Học Tập đã tỏ ra rất thành công, giúp giải quyết nhiều vấn đề thách thức trong công nghiệp và học thuật.

2.1 Thuật toán Máy Học Tập (machine learning - ML)

Thuật ngữ Máy Học Tập (ML) được đặt ra vào năm 1959 bởi Arthur Samuel, một nhân viên IBM và là người tiên phong trong lĩnh vực máy tính chơi game và trí tuệ nhân tạo. Vào đầu những năm 1960, một thử nghiệm Máy Học Tập với bộ nhớ băng đục lỗ, được gọi là Cybertron, đã được phát triển bởi Công ty Raytheon để phân tích tín hiệu sonar, điện tâm đồ và các mẫu giọng nói bằng cách sử dụng phương pháp học tăng cường thô sơ. Nó được người vận hành / giáo viên "huấn luyện" lặp đi lặp lại để nhận ra các mẫu và được trang bị một nút "goof" để khiến nó đánh giá lại các quyết định không chính xác. Một cuốn sách tiêu biểu về nghiên cứu máy học trong những năm 1960 là cuốn sách của Nilsson về Máy học, chủ yếu đề cập đến việc học máy để phân loại mẫu. Mối quan tâm liên quan đến nhận dạng mẫu tiếp tục vào những năm 1970, như được mô tả bởi Duda và Hart vào năm 1973. Năm 1981, một báo cáo đã được đưa ra về việc sử dụng các chiến lược giảng dạy để một mạng nơ-ron học cách nhận dạng 40 ký tự (26 chữ cái, 10 chữ số và 4 ký hiệu đặc biệt) từ một thiết bị đầu cuối máy tính.



Hình 5: Máy Học Tập (ML)

Tom M. Mitchell đã đưa ra một định nghĩa được trích dẫn rộng rãi, chính thức hơn về các thuật toán được nghiên cứu trong lĩnh vực Máy Học Tập: "Một chương trình máy tính được cho là học hỏi từ kinh nghiệm E đối với một số loại nhiệm vụ T và đo lường hiệu suất P nếu hiệu suất của nó tại các tác vụ trong T, được đo bằng P, cải thiện theo trải nghiệm E. "Định nghĩa này về các nhiệm vụ mà máy học có liên quan đưa ra một định nghĩa hoạt động cơ bản hơn là xác định lĩnh vực này theo thuật ngữ nhận thức. Điều này theo sau đề xuất của Alan Turing trong bài báo "Máy tính và trí thông minh" của ông, trong đó câu hỏi "Máy móc có thể suy nghĩ không?" được thay thế bằng câu hỏi "Liệu máy móc có thể làm những gì chúng ta (với tư cách là các thực thể tư duy) có thể làm được không?"

Máy Học Tập ngày nay có hai mục tiêu, một là phân loại dữ liệu dựa trên các mô hình đã được phát triển, mục đích khác là đưa ra dự đoán cho các kết quả trong tương lai dựa trên các mô hình này. Một thuật toán giả định cụ thể để phân loại dữ liệu có thể sử dụng khả năng nhìn của máy tính về nốt ruồi cùng với việc học có giám sát để huấn luyện nó phân loại nốt ruồi ung thư. Một thuật toán học máy cho giao dịch chứng khoán có thể thông báo cho nhà giao dịch về những dự đoán tiềm năng trong tương lai.

Như một nỗ lực khoa học, Máy Học Tập đã phát triển vượt ra ngoài nhiệm vụ của trí tuệ nhân tạo. Trong những ngày đầu của AI như một bộ môn học thuật, một số nhà nghiên cứu quan tâm đến việc để máy móc học hỏi từ dữ liệu. Họ đã cố gắng tiếp cận vấn đề bằng nhiều phương pháp biểu tượng khác nhau, cũng như cái mà sau đó được gọi là "mạng nơ-ron"; chúng hầu hết là các perceptron và các mô hình khác sau này được phát hiện là phát minh lại của các mô hình thống kê tuyến tính tổng quát hóa. Lý luận xác suất cũng được sử dụng, đặc biệt là trong chẩn đoán y tế tự động. Tuy nhiên, việc chú trọng ngày càng nhiều vào phương pháp tiếp cận dựa trên tri thức và logic đã gây ra rạn nứt giữa AI và Máy Học Tập. Các hệ thống xác suất đã bị cản trở bởi các vấn đề lý thuyết và thực tế về thu thập và biểu diễn dữ liệu. Đến năm 1980, các hệ thống chuyên gia đã thống trị AI, và thống kê không còn được ưa chuộng. Công việc về học tập dựa trên biểu tượng / kiến thức đã tiếp tục trong AI, dẫn đến lập trình logic quy nạp, nhưng những nghiên cứu về thống kê học nằm ngoài lĩnh vực AI vẫn tiếp tục trong các chủ đề nhận dạng mẫu (pattern recognition) và truy xuất thông tin (information retrieval). Nghiên cứu mạng nơ-ron đã bị AI và khoa học máy tính bỏ rơi cũng trong cùng thời gian. Hướng nghiên cứu mạng nơ-ron cũng được các nhà nghiên cứu bao gồm Hopfield, Rumelhart và Hinton tiếp tục bên ngoài lĩnh vực AI / CS, với tên gọi "connectionism". Thành công chính của họ đến vào giữa những năm 1980 với việc tái phát minh lại thuật toán Lan Truyền Ngược (backpropagation). Máy Học Tập (ML), được tổ chức lại thành một lĩnh vực riêng biệt, bắt đầu phát triển mạnh mẽ vào những năm 1990. Lĩnh vực này đã thay đổi mục tiêu từ đạt được trí thông minh nhân tạo sang giải quyết các vấn đề có thể giải quyết được có tính chất thực tế. Nó chuyển trọng tâm ra khỏi các phương pháp tiếp cận biểu tượng mà nó đã thừa hưởng từ AI, và sang các phương pháp và mô hình vay mượn từ thống kê, logic mờ và lý thuyết xác suất. Sự khác biệt giữa ML và AI thường bị hiểu nhầm. ML học hỏi và dự đoán dựa trên các quan sát thụ động, trong khi AI ngụ ý một tác nhân tương tác với môi trường để

tìm hiểu và thực hiện các hành động nhằm tối đa hóa cơ hội đạt được thành công mục tiêu của mình. Tính đến năm 2020, nhiều nguồn tiếp tục khẳng định rằng ML vẫn là một lĩnh vực con của AI. Những người khác có quan điểm rằng không phải tất cả ML đều là một phần của AI, mà chỉ một 'tập hợp con thông minh' của ML mới được coi là AI.

Các phương pháp tiếp cận máy học theo truyền thống được chia thành ba loại lớn, tùy thuộc vào bản chất của "tín hiệu" (signal) hoặc "phản hồi" (feedback) có sẵn cho hệ thống học tập.

Học có giám sát (Supervised learning): Các thuật toán học có giám sát xây dựng mô hình toán học của một tập hợp dữ liệu có chứa cả đầu vào và đầu ra mong muốn. Dữ liệu được gọi là dữ liệu đào tạo và bao gồm một tập hợp các ví dụ đào tạo. Mỗi ví dụ huấn luyện có một hoặc nhiều đầu vào và đầu ra mong muốn, còn được gọi là tín hiệu giám sát. Trong mô hình toán học, mỗi ví dụ huấn luyện được biểu diễn bằng một mảng hoặc vectơ, đôi khi được gọi là vectơ đặc trưng và dữ liệu huấn luyện được biểu diễn bằng ma trận. Thông qua tối ưu hóa lặp đi lặp lại của một chức năng mục tiêu, các thuật toán học có giám sát học một chức năng có thể được sử dụng để dự đoán đầu ra liên quan đến đầu vào mới. Một chức năng tối ưu sẽ cho phép thuật toán xác định chính xác đầu ra cho các đầu vào không phải là một phần của dữ liệu huấn luyện. Một thuật toán cải thiện độ chính xác của kết quả đầu ra hoặc dự đoán của nó theo thời gian được cho là đã học cách thực hiện nhiệm vụ đó. Các loại thuật toán học có giám sát bao gồm học tích cực (active), phân loại (classification) và hồi quy (regression). Thuật toán phân loại được sử dụng khi đầu ra bị giới hạn trong một tập giá trị giới hạn và thuật toán hồi quy được sử dụng khi đầu ra có thể có bất kỳ giá trị số nào trong một phạm vi. Ví dụ, đối với một thuật toán phân loại lọc email, đầu vào sẽ là một email đến và đầu ra sẽ là tên của thư mục chứa email. Học tương tự (Similarity learning) là một lĩnh vực của học máy có giám sát liên quan chặt chẽ đến hồi quy và phân loại, nhưng mục tiêu là học từ các ví dụ bằng cách sử dụng một hàm tương tự để đo lường mức độ tương tự hoặc liên quan của hai đối tượng. Nó có các ứng dụng trong hệ thống xếp hạng, đề xuất, theo dõi nhận dạng trực quan, xác minh khuôn mặt và xác minh người nói.

Học không giám sát (Unsupervised learning): Các thuật toán học không giám sát lấy một tập hợp dữ liệu chỉ chứa đầu vào và tìm cấu trúc trong dữ liệu, như nhóm hoặc phân cụm các điểm dữ liệu. Do đó, các thuật toán học hỏi từ dữ liệu thử nghiệm chưa được gắn nhãn, phân loại hoặc phân loại. Thay vì trả lời phản hồi, các thuật toán học tập không giám sát xác định những điểm chung trong dữ liệu và phản ứng dựa trên sự hiện diện hoặc vắng mặt của những điểm tương đồng đó trong mỗi phần dữ liệu mới. Một ứng dụng trung tâm của học không giám sát là trong lĩnh vực ước tính mật độ trong thống kê, chẳng hạn như tìm hàm mật độ xác suất. [37] Mặc dù học tập không giám sát bao gồm các lĩnh vực khác liên quan đến việc tóm tắt và giải thích các tính năng dữ liệu. Phân tích cụm là việc gán một tập hợp các quan sát thành các tập con (được gọi là các cụm) sao cho các quan sát trong cùng một cụm là tương tự nhau theo một hoặc nhiều tiêu chí được chỉ định trước, trong khi các quan sát được rút ra từ các cụm khác nhau là không giống nhau. Các kỹ thuật phân nhóm khác nhau đưa ra các giả

định khác nhau về cấu trúc của dữ liệu, thường được xác định bằng một số thước đo độ tương đồng và được đánh giá, ví dụ: theo độ gọn bên trong hoặc sự giống nhau giữa các thành viên của cùng một cụm và sự phân tách, sự khác biệt giữa các cụm. Các phương pháp khác dựa trên mật độ ước tính và kết nối đồ thị.

Học tập bán giám sát (Semi-supervised learning): Học tập bán giám sát nằm giữa học tập không giám sát (không có bất kỳ dữ liệu đào tạo nào được gắn nhãn) và học tập có giám sát (với dữ liệu đào tạo được gắn nhãn hoàn toàn). Một số ví dụ đào tạo bị thiếu nhãn đào tạo, nhưng nhiều nhà nghiên cứu học máy đã phát hiện ra rằng dữ liệu không được gắn nhãn, khi được sử dụng cùng với một lượng nhỏ dữ liệu được gắn nhãn, có thể tạo ra sự cải thiện đáng kể về độ chính xác của việc học. Trong quá trình học tập được giám sát lỏng lẻo, các nhãn đào tạo có chứa nhiễu, bị hạn chế hoặc không chính xác. Tuy nhiên, các nhãn này thường rẻ hơn để có được, dẫn đến các tập huấn luyện hiệu quả lớn hơn.

Học tăng cường (Reinforcement learning): Học tăng cường là một lĩnh vực của học máy liên quan đến cách các tác nhân phần mềm phải thực hiện hành động trong môi trường để tối đa hóa một số khái niệm về phần thưởng tích lũy. Do tính tổng quát của nó, lĩnh vực này được nghiên cứu trong nhiều ngành khác, chẳng hạn như lý thuyết trò chơi, lý thuyết điều khiển, nghiên cứu hoạt động, lý thuyết thông tin, tối ưu hóa dựa trên mô phỏng, hệ thống đa tác nhân, trí thông minh bầy đàn, thống kê và thuật toán di truyền. Trong học máy, môi trường thường được biểu diễn dưới dạng quy trình quyết định Markov (MDP). Nhiều thuật toán học tăng cường sử dụng kỹ thuật lập trình động. Các thuật toán học củng cố không giả định kiến thức về mô hình toán học chính xác của MDP và được sử dụng khi các mô hình chính xác không khả thi. Các thuật toán học tập củng cố được sử dụng trong các phương tiện tự hành hoặc học cách chơi trò chơi với đối thủ là con người.

Thực thi Máy Học Tập liên quan đến việc tạo ra một mô hình, được đào tạo trên một số dữ liệu đào tạo và sau đó có thể xử lý dữ liệu bổ sung để đưa ra dự đoán. Nhiều loại mô hình khác nhau đã được sử dụng và nghiên cứu cho các hệ thống Máy Học Tập.

Mạng nơ-ron nhân tạo: Mạng nơ-ron nhân tạo (ANN) hay còn gọi là hệ thống kết nối, là những hệ thống máy tính được lấy cảm hứng từ các mạng nơ-ron sinh học cấu thành não động vật một cách mơ hồ. Các hệ thống như vậy "học" để thực hiện các tác vụ bằng cách xem xét các ví dụ, nói chung mà không được lập trình với bất kỳ quy tắc cụ thể cho tác vụ nào. ANN là một mô hình dựa trên tập hợp các đơn vị hoặc nút được kết nối được gọi là "tế bào thần kinh nhân tạo", mô hình hóa lỏng lẻo các tế bào thần kinh trong não sinh học. Mỗi kết nối, giống như các khớp thần kinh trong bộ não sinh học, có thể truyền thông tin, một "tín hiệu", từ một tế bào thần kinh nhân tạo này sang một tế bào thần kinh nhân tạo khác. Một tế bào thần kinh nhân tạo nhận được tín hiệu có thể xử lý nó và sau đó phát tín hiệu cho các tế bào thần kinh nhân tạo bổ sung được kết nối với nó. Trong các triển khai ANN thông thường, tín hiệu tại kết nối giữa các nơ-ron nhân tạo là một số thực và đầu ra của mỗi nơ-ron nhân tạo được tính bằng một số hàm phi tuyến tính của tổng các đầu vào của nó. Các kết nối giữa các nơ-ron nhân tạo được gọi là "các cạnh". Các tế bào thần kinh và các cạnh nhân tạo

thường có trọng lượng điều chỉnh khi quá trình học tập diễn ra. Trọng lượng làm tăng hoặc giảm cường độ của tín hiệu tại một kết nối. Tế bào thần kinh nhân tạo có thể có một ngưỡng sao cho tín hiệu chỉ được gửi đi nếu tín hiệu tổng hợp vượt qua ngưỡng đó. Thông thường, các tế bào thần kinh nhân tạo được tập hợp lại thành từng lớp. Các lớp khác nhau có thể thực hiện các loại biến đổi khác nhau trên đầu vào của chúng. Tín hiệu đi từ lớp đầu tiên (lớp đầu vào) đến lớp cuối cùng (lớp đầu ra), có thể sau khi đi qua các lớp nhiều lần. Mạng nơ-ron nhân tạo đã được sử dụng cho nhiều nhiệm vụ khác nhau, bao gồm thị giác máy tính, nhận dạng giọng nói, dịch máy, lọc mạng xã hội, chơi board và trò chơi điện tử và chẩn đoán y tế.

Cây quyết định (Decision tree): Mô hình cây quyết định sử dụng cây quyết định như một mô hình dự đoán để đi từ quan sát về một mục (được biểu thị trong các nhánh) đến kết luận về giá trị mục tiêu của mục (được biểu thị trong các lá). Đây là một trong những cách tiếp cận mô hình dự đoán được sử dụng trong thống kê, khai thác dữ liệu và học máy. Mô hình cây trong đó biến mục tiêu có thể nhận một tập giá trị rời rạc được gọi là cây phân loại; trong các cấu trúc cây này, lá đại diện cho các nhãn lớp và các nhánh biểu thị các liên kết của các đối tượng địa lý dẫn đến các nhãn lớp đó. Cây quyết định trong đó biến mục tiêu có thể nhận các giá trị liên tục (thường là số thực) được gọi là cây hồi quy. Trong phân tích quyết định, cây quyết định có thể được sử dụng để thể hiện một cách trực quan và rõ ràng các quyết định và việc ra quyết định. Trong khai thác dữ liệu, cây quyết định mô tả dữ liệu, nhưng cây phân loại kết quả có thể là đầu vào để ra quyết định.

Support-vector machines: Support-vector machines (SVM), còn được gọi là mạng vectơ hỗ trợ, là một tập hợp các phương pháp học có giám sát liên quan được sử dụng để phân loại và hồi quy. Đưa ra một tập hợp các ví dụ đào tạo, mỗi ví dụ được đánh dấu là thuộc một trong hai loại, thuật toán đào tạo SVM xây dựng một mô hình dự đoán liệu một ví dụ mới có thuộc loại này hay loại kia hay không. [70] Thuật toán huấn luyện SVM là một bộ phân loại tuyến tính, nhị phân, không theo xác suất, mặc dù các phương pháp như chia tỷ lệ Platt tồn tại để sử dụng SVM trong cài đặt phân loại theo xác suất. Ngoài việc thực hiện phân loại tuyến tính, SVM có thể thực hiện phân loại phi tuyến tính một cách hiệu quả bằng cách sử dụng cái được gọi là thủ thuật hạt nhân, ánh xạ ngầm các đầu vào của chúng vào không gian đặc trưng chiều cao.

Phân tích hồi quy (regression analysis): Phân tích hồi quy bao gồm nhiều phương pháp thống kê để ước tính mối quan hệ giữa các biến đầu vào và các đặc trưng liên quan của chúng. Dạng phổ biến nhất của nó là hồi quy tuyến tính, trong đó một đường đơn được vẽ để phù hợp nhất với dữ liệu đã cho theo một tiêu chí toán học, chẳng hạn như bình phương nhỏ nhất thông thường. Phương pháp sau thường được mở rộng bằng các phương pháp chính quy hóa (toán học) để giảm thiểu việc trang bị quá mức và sai lệch, như trong hồi quy sườn núi. Khi giải quyết các vấn đề phi tuyến tính, các mô hình đi đến bao gồm hồi quy đa thức (ví dụ, được sử dụng để điều chỉnh đường xu hướng trong Microsoft Excel), hồi quy logistic (thường được sử dụng trong phân loại thống kê) hoặc thậm chí hồi

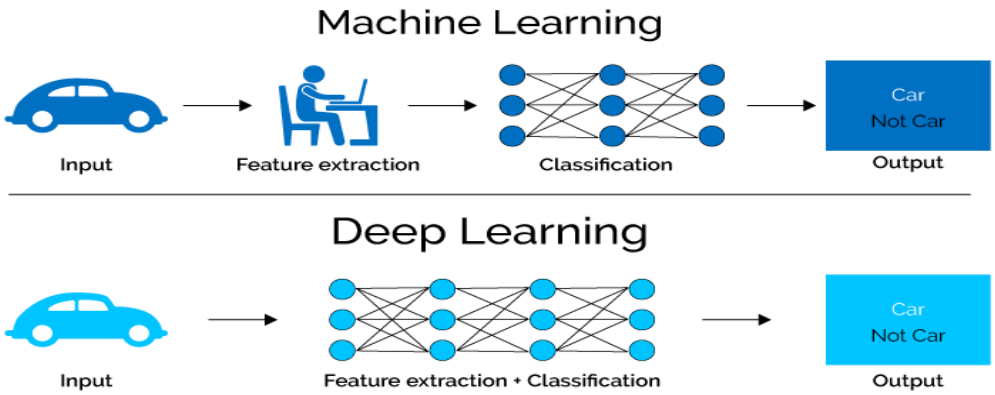
quy hạt nhân, giới thiệu tính phi tuyến tính bằng cách tận dụng thủ thuật hạt nhân để ánh xạ ngầm các biến đầu vào sang không gian chiều cao hơn.

Mạng Bayes (Bayesian networks): Mạng Bayes, mạng niềm tin hoặc mô hình đồ họa vòng có hướng là một mô hình đồ họa xác suất đại diện cho một tập hợp các biến ngẫu nhiên và sự độc lập có điều kiện của chúng bằng một đồ thị xoay chiều có hướng (DAG). Ví dụ, một mạng Bayes có thể đại diện cho các mối quan hệ xác suất giữa các bệnh và các triệu chứng. Với các triệu chứng, mạng có thể được sử dụng để tính toán xác suất của sự hiện diện của các bệnh khác nhau. Các thuật toán hiệu quả tồn tại để thực hiện suy luận và học hỏi. Mạng Bayes mô hình chuỗi các biến, như tín hiệu giọng nói hoặc chuỗi protein, được gọi là mạng Bayes động. Các khái quát của mạng Bayes có thể biểu diễn và giải quyết các vấn đề quyết định trong điều kiện không chắc chắn được gọi là sơ đồ ảnh hưởng.

Thông thường, các mô hình học máy yêu cầu một lượng lớn dữ liệu đáng tin cậy để các mô hình thực hiện các dự đoán chính xác. Khi đào tạo một mô hình học máy, các kỹ sư học máy cần nhắm mục tiêu và thu thập một lượng lớn dữ liệu mẫu đại diện. Dữ liệu từ tập huấn luyện có thể đa dạng như một kho văn bản, một tập hợp các hình ảnh, dữ liệu cảm biến và dữ liệu được thu thập từ những người dùng riêng lẻ của một dịch vụ. Quá khớp (overfitting) là điều cần chú ý khi đào tạo mô hình học máy. Các mô hình được đào tạo lấy từ dữ liệu sai lệch hoặc không được đánh giá có thể dẫn đến các dự đoán sai lệch hoặc không mong muốn. Mô hình thiên vị có thể dẫn đến các kết quả bất lợi do đó làm tăng thêm các tác động tiêu cực đến xã hội hoặc mục tiêu. Độ chệch của thuật toán là một kết quả tiềm ẩn từ dữ liệu không được chuẩn bị đầy đủ để đào tạo. Đạo đức học máy đang trở thành một lĩnh vực nghiên cứu và đáng chú ý là được tích hợp trong các nhóm kỹ sư học máy.

2.2 Thuật toán Học Sâu (deep learning - DL)

Học sâu (DL) là một lớp thuật toán Máy Học Tập sử dụng nhiều lớp cấu trúc mạng nơ-ron nhân tạo để trích xuất dần dần các đặc trưng cấp cao từ đầu vào thô. Ví dụ, trong xử lý hình ảnh, các lớp cấu trúc thấp hơn có thể xác định các cạnh, trong khi các lớp cấu trúc cao hơn có thể xác định các khái niệm liên quan đến con người như chữ số hoặc chữ cái hoặc khuôn mặt.



Hình 6: Điểm khác nhau giữa Máy Học Tập và Học Sâu

Hầu hết các mô hình học sâu hiện đại đều dựa trên mạng nơ-ron nhân tạo, cụ thể là mạng nơ-ron tích chập (Convolutional Neural Network - CNN), mặc dù chúng cũng có thể bao gồm các công thức mệnh đề hoặc các biến tiềm ẩn được tổ chức theo lớp trong các mô hình sinh trưởng sâu như các nút trong mạng deep belief networks và deep Boltzmann machines. Trong Học Sâu, mỗi cấp độ học cách chuyển đổi dữ liệu đầu vào của mình thành một biểu diễn tổng hợp và trừu tượng hơn một chút. Trong một ứng dụng nhận dạng hình ảnh, đầu vào thô có thể là một ma trận các pixel; lớp biểu diễn đầu tiên có thể trừu tượng hóa các pixel và mã hóa các cạnh; lớp thứ hai có thể soạn và mã hóa sự sắp xếp của các cạnh; lớp thứ ba có thể mã hóa mũi và mắt; và lớp thứ tư có thể nhận ra rằng hình ảnh có một khuôn mặt. Quan trọng là, quá trình học sâu có thể tự mình tìm hiểu các tính năng nào cần đặt ở cấp độ nào một cách tối ưu. Điều này không loại bỏ nhu cầu điều chỉnh bằng tay; ví dụ, số lượng lớp và kích thước lớp khác nhau có thể cung cấp các mức độ trừu tượng khác nhau. Từ "sâu" trong "học sâu" dùng để chỉ số lớp mà dữ liệu được chuyển đổi qua đó. Chính xác hơn, các hệ thống học tập sâu có chiều sâu đáng kể về lộ trình phân bổ tín chỉ (CAP). CAP là chuỗi chuyển đổi từ đầu vào đến đầu ra. CAP mô tả các kết nối nhân quả có thể xảy ra giữa đầu vào và đầu ra. Đối với mạng nơron truyền thẳng, độ sâu của các CAP là của mạng và là số lớp ẩn cộng với một (vì lớp đầu ra cũng được tham số hóa). Đối với mạng nơ-ron tuần hoàn, trong đó một tín hiệu có thể truyền qua một lớp nhiều hơn một lần, độ sâu CAP có khả năng không giới hạn. Không có ngưỡng độ sâu được thống nhất rộng rãi nào phân chia học nông với học sâu, nhưng hầu hết các nhà nghiên cứu đều đồng ý rằng học sâu liên quan đến độ sâu CAP cao hơn 2. CAP độ sâu 2 đã được chứng minh là một đại lượng xấp xỉ phổ quát theo nghĩa là nó có thể mô phỏng bất kỳ chức năng nào. Ngoài ra, nhiều lớp hơn không làm tăng thêm khả năng xấp xỉ hàm của mạng. Mô hình sâu (CAP> 2) có thể trích xuất các tính năng tốt hơn mô hình nông và do đó, các lớp bổ sung giúp học các tính năng một cách hiệu quả. Kiến trúc học sâu có thể được xây dựng bằng phương pháp từng lớp một tham lam. Học sâu giúp gỡ rối những điều trừu tượng này và chọn ra những tính năng nào cải thiện hiệu suất.

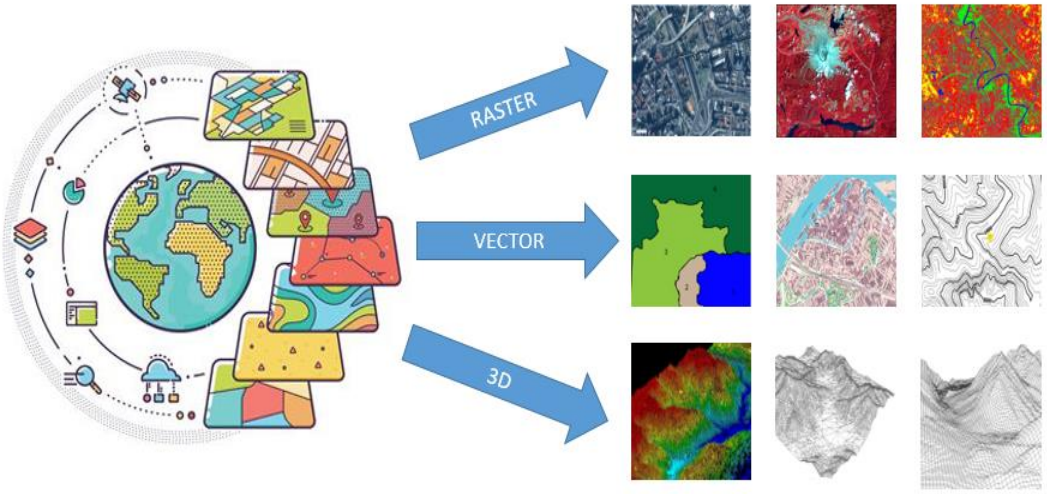
3. Các bài toán cơ bản khi làm việc với dữ liệu không gian địa lý

3.1 Dữ liệu không gian địa lý (geospatial data)

Dữ liệu không gian địa lý là thông tin mô tả các đối tượng, sự kiện hoặc các đối tượng địa lý khác có vị trí trên hoặc gần bề mặt trái đất. Dữ liệu không gian địa lý thường kết hợp thông tin vị trí (thường là tọa độ trên trái đất) và thông tin thuộc tính (đặc điểm của đối tượng, sự kiện hoặc hiện tượng liên quan) với thông tin thời gian (thời gian hoặc tuổi thọ mà vị trí và thuộc tính tồn tại). Vị trí được cung cấp có thể là tĩnh trong thời gian ngắn (ví dụ: vị trí của một thiết bị, một sự kiện động đất, trẻ em sống trong cảnh nghèo đói) hoặc động (ví dụ, một phương tiện di chuyển hoặc người đi bộ, sự lây lan của bệnh truyền nhiễm) .

Dữ liệu không gian địa lý thường liên quan đến tập hợp lớn dữ liệu không gian được thu thập từ nhiều nguồn đa dạng với các định dạng khác nhau và có thể bao gồm thông tin như dữ liệu điều tra dân số, hình ảnh vệ tinh, dữ liệu thời

tiết, dữ liệu điện thoại di động, hình ảnh được vẽ và dữ liệu truyền thông xã hội. Dữ liệu không gian địa lý hữu ích nhất khi nó có thể được khám phá, chia sẻ, phân tích và sử dụng kết hợp với dữ liệu kinh doanh truyền thống.



Hình 7: Dữ liệu không gian địa lý

Phân tích không gian địa lý (Geospatial analytics) được sử dụng để thêm thời gian và vị trí vào các loại dữ liệu truyền thống và để xây dựng hình ảnh trực quan hóa dữ liệu. Những hình ảnh trực quan này có thể bao gồm bản đồ, biểu đồ, số liệu thống kê và bản đồ thể hiện những thay đổi trong lịch sử và những thay đổi hiện tại. Bối cảnh bổ sung này cho phép có một bức tranh toàn cảnh hơn về các sự kiện. Những thông tin chi tiết có thể bị bỏ qua trong một bảng tính lớn được hiển thị dưới dạng hình ảnh và mẫu trực quan dễ nhận ra. Điều này có thể làm cho các dự đoán nhanh hơn, dễ dàng hơn và chính xác hơn.

Hệ thống thông tin không gian địa lý (GIS) liên quan cụ thể đến việc lập bản đồ vật lý của dữ liệu trong một biểu diễn trực quan. Ví dụ: khi bản đồ bão (hiển thị vị trí và thời gian) được phủ lên bởi một lớp khác hiển thị các khu vực có khả năng xảy ra sét đánh, bạn sẽ thấy GIS đang hoạt động.

Dữ liệu không gian địa lý là thông tin được ghi lại cùng với một chỉ báo địa lý của một số loại. Có hai dạng dữ liệu không gian địa lý chính: dữ liệu vectơ và dữ liệu raster. Dữ liệu vectơ là dữ liệu trong đó các điểm, đường và đa giác đại diện cho các đối tượng địa lý như thuộc tính, thành phố, đường, núi và vùng nước. Ví dụ: một biểu diễn trực quan sử dụng dữ liệu vectơ có thể bao gồm các ngôi nhà được biểu thị bằng điểm, đường được biểu thị bằng đường và toàn bộ thị trấn được biểu thị bằng đa giác. Dữ liệu raster là các ô được phân chia theo pixel hoặc ô lưới được xác định theo hàng và cột. Dữ liệu raster tạo ra hình ảnh về cơ bản phức tạp hơn, chẳng hạn như ảnh chụp và ảnh vệ tinh.

Geospatial big data: Đối phó với các tập dữ liệu không gian địa lý lớn đặt ra nhiều thách thức. Vì lý do này, nhiều tổ chức đấu tranh để tận dụng tối đa dữ liệu không gian địa lý. Đầu tiên, đó là khối lượng dữ liệu không gian địa lý tuyệt đối. Ví dụ, người ta ước tính rằng 100 TB dữ liệu liên quan đến thời tiết được tạo ra hàng ngày. Chỉ riêng điều này đã gây ra các vấn đề về lưu trữ và truy cập

đáng kể cho hầu hết các tổ chức. Dữ liệu không gian địa lý cũng được lưu trữ trên nhiều tệp khác nhau, điều này gây khó khăn cho việc tìm tệp chứa dữ liệu cần thiết để giải quyết vấn đề cụ thể của bạn. Ngoài ra, dữ liệu không gian địa lý được lưu trữ ở nhiều định dạng khác nhau và được hiệu chỉnh theo các tiêu chuẩn khác nhau. Bất kỳ nỗ lực nào để so sánh, kết hợp hoặc lập bản đồ dữ liệu trước tiên đều đòi hỏi một lượng dữ liệu đáng kể và định dạng lại. Cuối cùng, làm việc với dữ liệu không gian địa lý thô đòi hỏi kiến thức chuyên môn và ứng dụng toán học nâng cao để tiến hành các tác vụ cần thiết, chẳng hạn như căn chỉnh không gian địa lý của các lớp dữ liệu. Trừ khi các nhà phân tích thành thạo và có kinh nghiệm trong công việc này, họ sẽ không nhận được giá trị từ dữ liệu hoặc không đạt được tiến bộ đối với các mục tiêu kinh doanh của tổ chức họ.

Geospatial big data collection: Bởi vì khối lượng dữ liệu không gian địa lý mà các doanh nghiệp yêu cầu thường xuyên là rất lớn, nhiều tổ chức tìm cách sử dụng dịch vụ để có được dữ liệu không gian địa lý được quản lý. Bất kể bạn lấy nguồn dữ liệu không gian địa lý của mình từ đâu, chất lượng dữ liệu phải luôn được duy trì. Dữ liệu kém dẫn đến các mô hình được sử dụng ít hoặc hạn chế. (Cụm từ cảnh báo "Dữ liệu xấu trong - thông tin chi tiết xấu" chứng minh sự thật một cách tàn bạo.) Có vẻ như các tổ chức có thể được hưởng lợi đáng kể từ việc có một giải pháp tại chỗ để quản lý và kiểm tra dữ liệu, vì vậy mọi dữ liệu "rác" đều được tính đúng cách.

Geospatial big data management: Với rất nhiều dữ liệu hiện đang dồi dào, việc quản lý nó có tầm quan trọng đáng kể. Nhiều tổ chức nhận thấy dữ liệu bị tràn ngập và đang chuyển sang các nhà khoa học dữ liệu nội bộ của họ để giúp họ quản lý nó. Người ta ước tính rằng có tới 90% thời gian của các nhà khoa học dữ liệu dành cho các hoạt động quản lý dữ liệu, bao gồm tổ chức, "làm sạch" và định dạng lại dữ liệu. Điều đó khiến các nhà khoa học dữ liệu chỉ có 10% thời gian trong ngày làm việc của họ để dành cho việc phân tích xu hướng dữ liệu và sử dụng những thông tin chi tiết đó để giúp hình thành chính sách kinh doanh.

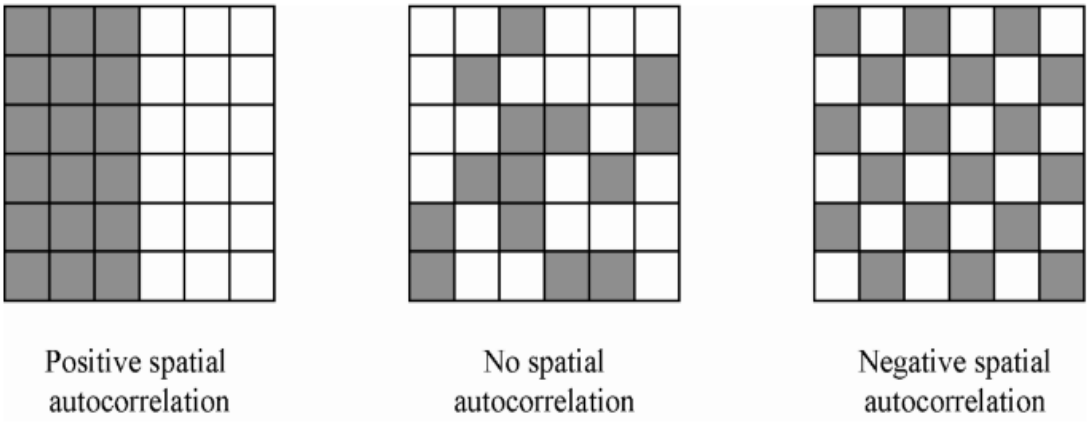
Cả hoạt động thu thập dữ liệu và quản lý dữ liệu có thể được thực thi hiệu quả hơn. Giải pháp có thể mở rộng, dựa trên đám mây và có thể chứa các định dạng tệp khác nhau. Bằng cách sử dụng cơ sở dữ liệu thông tin được tối ưu hóa được quản lý, các nhà khoa học dữ liệu có thể có nhiều thời gian hơn để tập trung vào cách sử dụng thông tin chi tiết phân tích và chuyển đổi chúng thành tiến trình của tổ chức và tác động của doanh nghiệp.

3.2 Bài toán khai phá dữ liệu không gian (ESDA)

Trong Khoa học dữ liệu, chúng tôi có xu hướng khám phá và điều tra dữ liệu trước khi thực hiện bất kỳ tác vụ xử lý hoặc mô hình hóa nào. Điều này giúp bạn xác định các mẫu, tóm tắt các đặc điểm chính của dữ liệu hoặc kiểm tra giả thuyết. Phân tích dữ liệu Khám phá thông thường không điều tra thành phần vị trí của tập dữ liệu một cách rõ ràng mà thay vào đó xử lý mối quan hệ giữa các biến và cách chúng ảnh hưởng đến nhau. Phương pháp thống kê tương quan thường được sử dụng để khám phá mối quan hệ giữa các biến.

Ngược lại, khai phá dữ liệu không gian (ESDA) tương quan một biến cụ thể với một vị trí, có tính đến các giá trị của cùng một biến trong vùng lân cận.

Các phương pháp được sử dụng cho mục đích này được gọi là tự tương quan không gian (Spatial Autocorrelation - SA). Tự tương quan không gian là mô tả sự hiện diện (hoặc vắng mặt) của các biến thể không gian trong một biến nhất định. Giống như các phương pháp tương quan thông thường, Tự tương quan không gian có giá trị âm và dương. Tự tương quan không gian dương là khi các khu vực gần nhau có các giá trị tương tự nhau (Cao-cao hoặc Thấp-thấp). Mặt khác, tự tương quan không gian âm chỉ ra rằng các khu vực lân cận khác nhau (Giá trị thấp bên cạnh giá trị cao).



Hình 8: Tự tương quan không gian

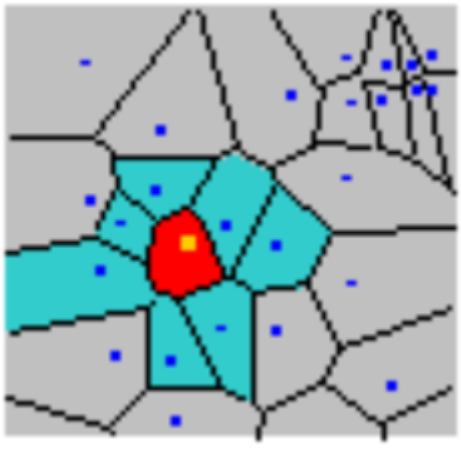
Phân phối của dữ liệu (distribution of data): Hầu hết các phương pháp nội suy được cung cấp bởi Geostatistical Analyst không yêu cầu dữ liệu phải được phân phối chuẩn, mặc dù trong trường hợp này, bản đồ dự đoán có thể không tối ưu. Tuy nhiên, một số phương pháp kriging nhất định yêu cầu dữ liệu phải được phân phối gần như bình thường (gần với đường cong hình chuông). Đặc biệt, các bản đồ lượng tử và xác suất được tạo bằng cách sử dụng kriging thông thường, đơn giản hoặc phổ quát giả định rằng dữ liệu đến từ phân phối chuẩn đa biến. Ngoài ra, các mô hình kriging đơn giản, được sử dụng làm cơ sở cho mô phỏng thống kê địa lý (xem Mô phỏng thống kê địa lý Gauss để biết thêm thông tin) nên sử dụng dữ liệu được phân phối bình thường hoặc bao gồm phép biến đổi điểm bình thường như một phần của mô hình để đảm bảo điều này. Tất cả các phương pháp kriging đều dựa trên giả định về tính ổn định. Giả định này một phần yêu cầu tất cả các giá trị dữ liệu đến từ các bản phân phối có cùng độ biến thiên. Các phép biến đổi dữ liệu cũng có thể được sử dụng để thỏa mãn giả định về độ biến thiên ngang nhau này.

Tìm kiếm các ngoại lệ toàn cục và địa phương: Ngoại lệ toàn cầu là một điểm mẫu được đo có giá trị rất cao hoặc rất thấp so với tất cả các giá trị trong tập dữ liệu. Ví dụ: nếu 99 trong số 100 điểm có giá trị từ 300 đến 400, nhưng điểm thứ 100 có giá trị là 750, thì điểm thứ 100 có thể là ngoại lệ toàn cầu. Ngoại lệ cục bộ là một điểm mẫu được đo có giá trị trong phạm vi bình thường cho toàn bộ tập dữ liệu, nhưng nếu bạn nhìn vào các điểm xung quanh, nó cao hoặc thấp bất thường. Ví dụ, sơ đồ dưới đây là một mặt cắt ngang của một thung lũng trong

một cảnh quan. Tuy nhiên, có một điểm ở trung tâm thung lũng có giá trị cao bất thường so với môi trường xung quanh nó, nhưng nó không có gì lạ so với toàn bộ tập dữ liệu. Điều quan trọng là phải xác định các ngoại lệ vì hai lý do: chúng có thể là những bất thường thực sự trong hiện tượng, hoặc giá trị có thể đã được đo lường hoặc ghi lại không chính xác. Nếu một ngoại lệ là một bất thường thực tế trong hiện tượng, thì đây có thể là điểm quan trọng nhất của nghiên cứu và để hiểu được hiện tượng. Ví dụ, một mẫu trên mạch của quặng khoáng sản có thể là ngoại lai và vị trí quan trọng nhất đối với một công ty khai thác. Nếu các lỗi ngoại lệ do lỗi trong quá trình nhập dữ liệu không chính xác rõ ràng, chúng nên được sửa hoặc loại bỏ trước khi tạo bề mặt. Các giá trị ngoại lai có thể có một số tác động bất lợi trên bề mặt dự đoán của bạn vì ảnh hưởng đến mô hình bán biểu đồ và ảnh hưởng của các giá trị lân cận.

Tìm kiếm các xu hướng toàn cục: Để xác định xu hướng toàn cầu trong dữ liệu của bạn, hãy tìm một đường cong không bằng phẳng trên mặt phẳng hình chiếu. Nếu bạn có xu hướng toàn cục trong dữ liệu của mình, bạn có thể muốn tạo bề mặt bằng cách sử dụng một trong các phương pháp nội suy xác định (ví dụ: đa thức tổng thể hoặc cục bộ) hoặc bạn có thể muốn loại bỏ xu hướng khi sử dụng kriging.

Kiểm tra biến động cục bộ: Đa giác Voronoi được tạo để mọi vị trí trong một đa giác đều gần với điểm mẫu trong đa giác đó hơn bất kỳ điểm mẫu nào khác. Sau khi các đa giác được tạo, các vùng lân cận của một điểm mẫu được xác định là bất kỳ điểm mẫu nào khác mà đa giác có chung đường viền với điểm mẫu đã chọn. Ví dụ, trong hình sau, điểm mẫu màu xanh lục sáng được bao quanh bởi một đa giác, được đánh dấu bằng màu đỏ. Mọi vị trí trong đa giác màu đỏ đều gần với điểm mẫu màu xanh lục sáng hơn bất kỳ điểm mẫu nào khác (được cho dưới dạng các chấm nhỏ màu xanh lam đậm). Các đa giác màu xanh lam đều có chung đường viền với đa giác màu đỏ, vì vậy các điểm mẫu trong các đa giác màu xanh lam là các điểm mẫu lân cận của điểm mẫu màu xanh lục sáng.



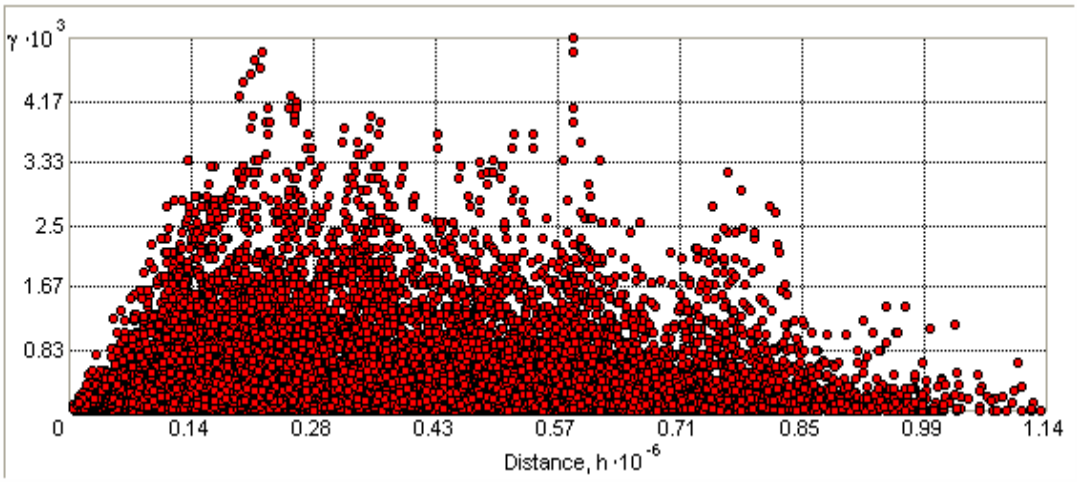
Hình 9: Đa giác Voronoi

Sử dụng định nghĩa này về các nước láng giềng, có thể tính toán nhiều số liệu thống kê địa phương. Ví dụ: giá trị trung bình cục bộ được tính bằng cách lấy giá trị trung bình của các điểm mẫu trong đa giác màu đỏ và xanh lam. Giá trị trung bình này sau đó được gán cho đa giác màu đỏ. Quá trình này được lặp lại cho tất cả các đa giác và các hình lân cận của chúng, và kết quả được hiển thị

bằng cách sử dụng đường dốc màu để giúp hình dung các vùng có giá trị cục bộ cao và thấp.Số liệu thống kê Voronoi có thể được sử dụng cho các mục đích khác nhau và có thể được nhóm thành các loại chức năng chung sau:

Functional category	Voronoi statistics
Local Smoothing	Mean, Mode, Median
Local Variation	Standard deviation, Interquartile range, Entropy
Local Outliers	Cluster
Local Influence	Simple

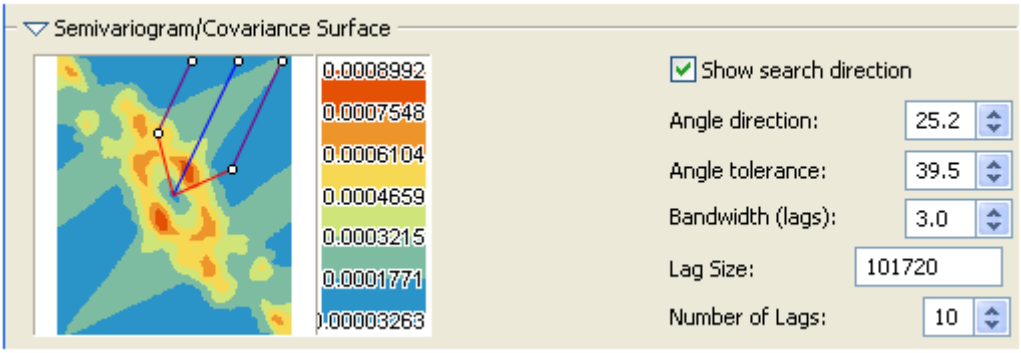
Kiểm tra tự tương quan không gian: Bằng cách khám phá dữ liệu của mình, bạn sẽ hiểu rõ hơn về tự tương quan không gian giữa các giá trị đo được. Sự hiểu biết này có thể được sử dụng để đưa ra quyết định tốt hơn khi lựa chọn các mô hình để dự đoán không gian. Công cụ Semivariogram / Covariance Cloud hiển thị các giá trị hiệp phương sai và bán biểu đồ thực nghiệm cho tất cả các cặp vị trí trong tập dữ liệu và vẽ biểu đồ chúng dưới dạng hàm của khoảng cách phân tách hai vị trí. Công cụ Semivariogram / Covariance Cloud có thể được sử dụng để kiểm tra các đặc điểm cục bộ của tự tương quan không gian trong tập dữ liệu và tìm kiếm các giá trị ngoại lai cục bộ. Đám mây bán biểu đồ trông giống như sau:



Hình 10: Semivariogram / Covariance Cloud

Trong hình minh họa ở trên, mỗi chấm đỏ hiển thị giá trị biểu đồ bán biến thực nghiệm (chênh lệch bình phương giữa các giá trị của hai điểm dữ liệu tạo thành một cặp) được vẽ dựa trên khoảng cách phân tách hai điểm. Mỗi điểm trong đám mây đại diện cho một cặp điểm trong tập dữ liệu, vì vậy số điểm trong đám mây sẽ tăng nhanh khi số điểm trong tập dữ liệu tăng lên. Đối với n điểm trong tập dữ liệu, đám mây hiệp phương sai / hiệp phương sai sẽ hiển thị n * (n-1) / 2 điểm. Vì lý do này, không nên sử dụng bộ dữ liệu có hơn vài nghìn điểm. Nếu tập dữ liệu của bạn chứa nhiều hơn một vài nghìn điểm, bạn nên sử dụng công cụ Tính năng tập hợp con để lấy mẫu ngẫu nhiên của các điểm và sử dụng tập hợp con

này trong đám mây biểu đồ bán biến / hiệp phương sai. Bề mặt biểu đồ bán nguyệt có khả năng Hướng tìm kiếm được hiển thị bên dưới. Các giá trị trong đám mây bán biểu đồ được đưa vào các thùng dựa trên hướng và khoảng cách giữa một cặp vị trí. Sau đó, các giá trị binned này được tính trung bình và làm mịn để tạo ra bề mặt bán biểu đồ. Trong hình bên dưới, chú giải hiển thị các giá trị giữa các quá trình chuyển đổi màu sắc. Trong công cụ này, bạn có thể nhập kích thước trễ để kiểm soát kích thước của thùng và số lượng thùng được xác định bởi số độ trễ mà bạn chỉ định. Mức độ của bề mặt bán biểu đồ được kiểm soát bởi kích thước độ trễ và số lượng độ trễ.



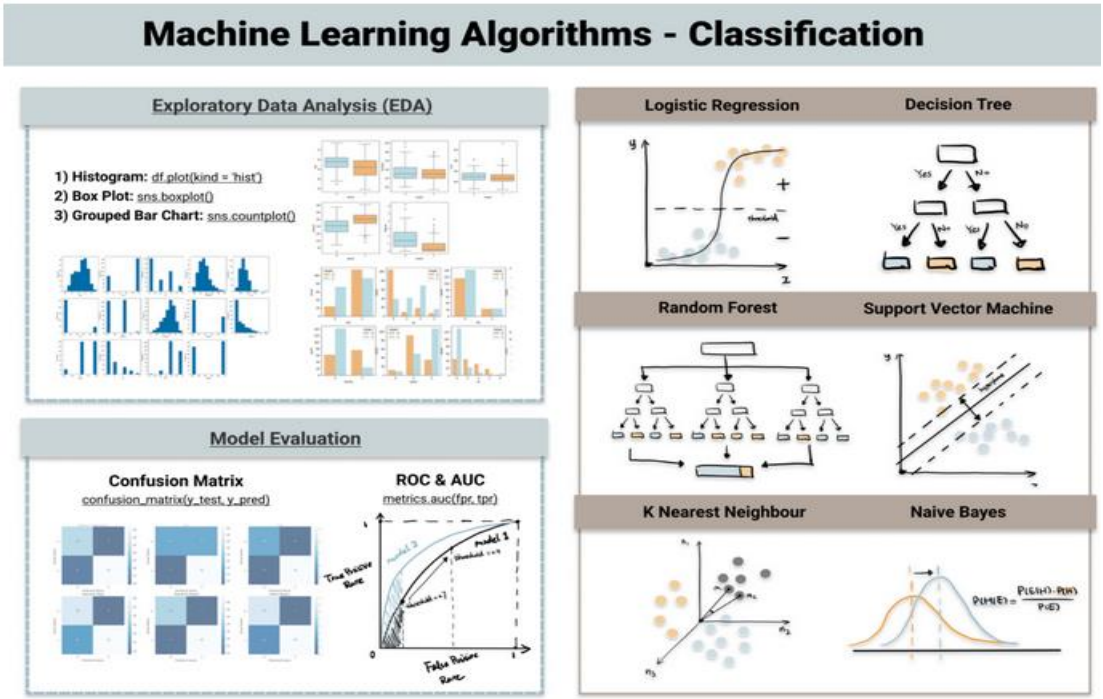
Hình 11: phát hiện hướng biến động

Không phải tất cả các bước này đều cần thiết trong mọi trường hợp. Ví dụ: nếu bạn quyết định sử dụng phương pháp nội suy không yêu cầu thước đo tự tương quan không gian (GPI, LPI hoặc RBF), thì không cần thiết phải khám phá tự tương quan không gian trong dữ liệu. Tuy nhiên, có thể là một ý tưởng hay để khám phá nó, vì một lượng lớn tự tương quan không gian có thể dẫn đến việc sử dụng một phương pháp nội suy khác (ví dụ: kriging) so với phương pháp bạn định sử dụng ban đầu.

3.3 Bài toán phân loại (classification)

Một số đặc điểm tiêu biểu của các hiện tượng không gian địa lý và dữ liệu môi trường: tính phi tuyến (các mô hình tuyến tính có khả năng ứng dụng hạn chế); sự không ổn định theo không gian và thời gian, tức là trong nhiều trường hợp không thể chấp nhận các giả thuyết về sự đứng yên trong không gian-thời gian (tính đứng yên bậc hai, các giả thuyết nội tại); sự thay đổi nhiều quy mô (sự thay đổi cao ở một số quy mô địa lý), sự hiện diện của tiếng ồn và các điểm cực đoan / ngoại lai; tính chất đa biến. Những "đặc điểm" này vi phạm các ứng dụng của các phương pháp truyền thống (bao gồm nhiều mô hình thống kê địa lý) và làm phức tạp hóa việc phân tích, mô hình hóa và trực quan hóa dữ liệu địa lý và môi trường. Như đã đề cập ở trên, trong nhiều tình huống thực tế, các vấn đề phải được xem xét trong các không gian đối tượng địa lý có chiều cao (đặc điểm địa lý) (rất thường kích thước của không gian này có thể lớn hơn 10). Nó bao gồm không gian địa lý gốc và nhiều đặc điểm có được từ các mô hình dựa trên khoa học hoặc các nguồn thông tin bổ sung, ví dụ, ảnh viễn thám; độ dốc, độ cong, v.v ... có nguồn gốc từ các mô hình độ cao kỹ thuật số. Trong trường hợp thứ hai, các mô hình thống kê (địa lý) truyền thống quá phức tạp để áp dụng hoặc không thể áp dụng chúng. Ví dụ, biến thể có thể được áp dụng hiệu quả trong

không gian có kích thước nhỏ hơn 3. Do đó, các câu hỏi quan trọng của phân tích và mô hình dữ liệu không gian và không gian - thời gian nói chung (bao gồm cả dự đoán và dự báo) đối phó với sự phát triển và triển khai dữ liệu -các mô hình đa biến, phi tuyến, mạnh mẽ và đa biến làm việc trong không gian chiều cao và có đặc tính tổng quát hóa tốt.



Hình 12: Các mô hình Máy Học Tập cho bài toán phân loại

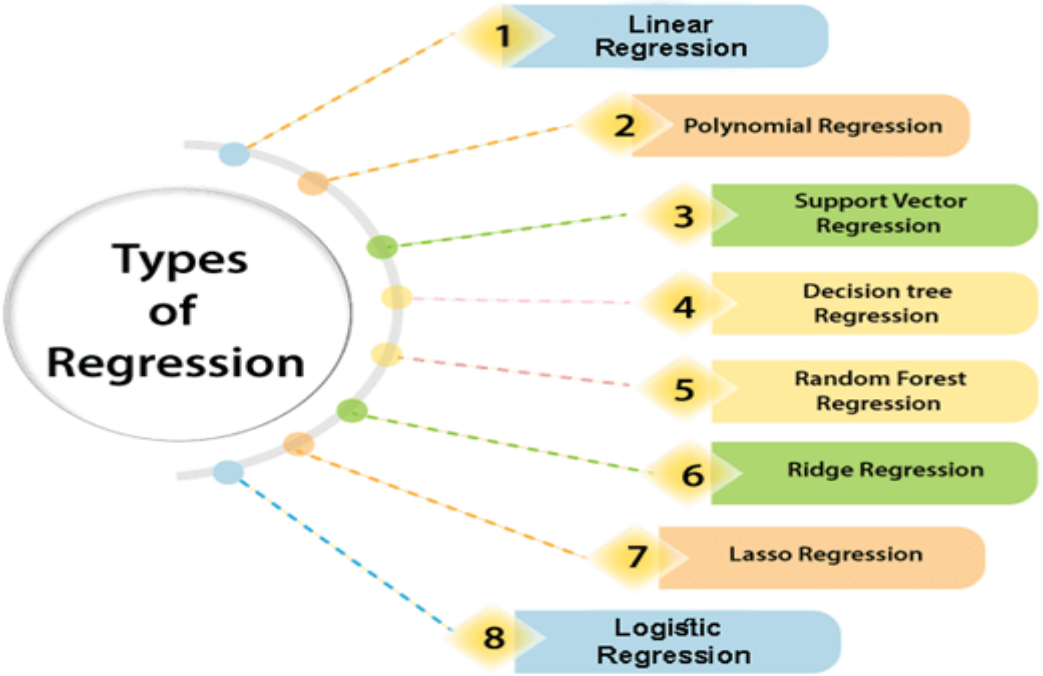
Một trong những giải pháp khả thi có thể dựa trên các thuật toán học máy, cụ thể là Mạng thần kinh nhân tạo của các kiến trúc khác nhau và Lý thuyết học thống kê (ví dụ: các phương pháp dựa trên hạt nhân: Máy vectơ hỗ trợ, Hồi quy vectơ hỗ trợ, v.v.). Chúng ta hãy đề cập rằng, các phương pháp tiếp cận như vậy, trở thành một hướng dữ liệu (hộp đen / xám) phụ thuộc rất nhiều vào chất lượng và số lượng dữ liệu. Do đó, việc áp dụng các công cụ thống kê / địa lý khác nhau để kiểm soát chất lượng phân tích và mô hình hóa dữ liệu bằng ML là hữu ích và cần thiết. Ví dụ, biến thể giúp hiểu và lập mô hình các tương quan dị hướng trong không gian, xu hướng không gian, biến thiên cục bộ và mức độ tiếng ồn

3.4 Bài toán hồi quy (regression)

Hồi quy (và dự đoán tổng quát hơn) cung cấp cho chúng ta một trường hợp hoàn hảo để kiểm tra cách cấu trúc không gian có thể giúp chúng ta hiểu và phân tích dữ liệu của mình. Trong chương này, chúng ta thảo luận về cách cấu trúc không gian có thể được sử dụng để xác nhận và cải thiện các thuật toán dự đoán, tập trung vào hồi quy tuyến tính một cách cụ thể.

Thông thường, cấu trúc không gian giúp mô hình hồi quy theo một trong hai cách. Cách đầu tiên (và rõ ràng nhất) không gian có thể có tác động đến dữ liệu của chúng ta là khi quá trình tạo dữ liệu chính là không gian rõ ràng. Ở đây, hãy nghĩ về một cái gì đó giống như giá cho những ngôi nhà dành cho một gia

đình. Thông thường, các cá nhân phải trả một khoản cao hơn giá nhà của họ để được sống trong một khu học chánh tốt hơn với cùng một ngôi nhà chất lượng. Ngoài ra, những ngôi nhà gần những nơi gây ô nhiễm tiếng ồn hoặc hóa chất như nhà máy xử lý nước thải, cơ sở tái chế hoặc đường cao tốc rộng, thực sự có thể rẻ hơn chúng ta dự đoán. Trong các trường hợp như mắc bệnh hen suyễn, các địa điểm mà các cá nhân có xu hướng đi lại trong ngày, chẳng hạn như nơi làm việc hoặc giải trí, có thể có nhiều ảnh hưởng đến sức khỏe của họ hơn là địa chỉ cư trú của họ. Trong trường hợp này, có thể cần sử dụng dữ liệu từ các địa điểm khác để dự đoán tỷ lệ mắc bệnh hen suyễn tại một địa điểm nhất định. Bất kể trường hợp cụ thể đang diễn ra là gì, ở đây, địa lý là một tính năng: nó trực tiếp giúp chúng ta đưa ra dự đoán về kết quả bởi vì những kết quả đó thu được từ các quá trình địa lý.



Hình 13: Các loại bài toán hồi quy

Một sự thay thế (và sự hiểu biết hoài nghi hơn) miễn cưỡng thừa nhận giá trị công cụ của địa lý. Thông thường, trong phân tích các phương pháp dự đoán và phân loại, chúng tôi quan tâm đến việc phân tích những gì chúng tôi nhận được sai. Điều này là phổ biến trong kinh tế lượng; một nhà phân tích có thể lo ngại rằng mô hình dự đoán sai một số loại quan sát có hệ thống. Nếu chúng tôi biết mô hình của mình thường hoạt động kém trên một tập hợp các quan sát hoặc loại đầu vào đã biết, chúng tôi có thể tạo ra một mô hình tốt hơn nếu chúng tôi có thể giải thích được điều này. Trong số các loại chẩn đoán lỗi khác, vị trí địa lý cung cấp cho chúng tôi một phương pháp nhúng đặc biệt hữu ích để đánh giá cấu trúc trong các lỗi của chúng tôi. Lỗi phân loại / dự đoán ánh xạ có thể giúp hiển thị có hay không có các cụm lỗi trong dữ liệu của chúng tôi. Nếu chúng ta biết rằng sai số có xu hướng lớn hơn ở một số khu vực so với các khu vực khác

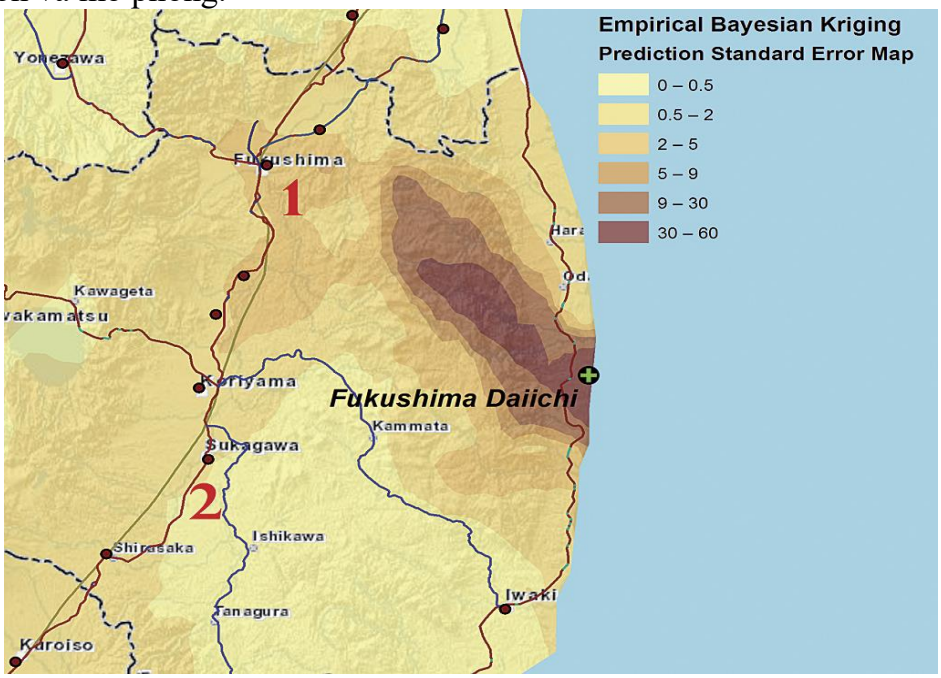
(hoặc nếu lỗi "lây lan" giữa các lần quan sát), thì chúng tôi có thể khai thác cấu trúc này để đưa ra dự đoán tốt hơn.

Lỗi cấu trúc không gian của chúng tôi có thể phát sinh từ khi địa lý nên là một thuộc tính bằng cách nào đó, nhưng chúng tôi không chắc chắn chính xác cách đưa nó vào mô hình của mình. Chúng cũng có thể phát sinh bởi vì có một số tính năng khác mà sự thiếu sót của chúng gây ra lỗi không gian mà chúng tôi thấy; nếu tính năng bổ sung này được bao gồm, cấu trúc sẽ biến mất. Hoặc, nó có thể phát sinh từ các tương tác phức tạp và phụ thuộc lẫn nhau giữa các đối tượng địa lý mà chúng tôi đã chọn để sử dụng làm yếu tố dự đoán, dẫn đến cấu trúc nội tại dự đoán sai. Hầu hết các yếu tố dự đoán mà chúng tôi sử dụng trong các mô hình về quá trình xã hội đều chứa thông tin không gian được thể hiện: tạo khuôn mẫu nội tại cho đặc điểm mà chúng tôi nhận được miễn phí trong mô hình. Nếu chúng ta có ý định hoặc không, việc sử dụng công cụ dự đoán theo mô hình không gian trong một mô hình có thể dẫn đến lỗi theo mô hình không gian; sử dụng nhiều hơn một có thể khuếch đại hiệu ứng này. Do đó, bất kể quá trình thực sự có mang tính địa lý rõ ràng hay không, thông tin bổ sung về mối quan hệ không gian giữa các quan sát của chúng tôi hoặc thông tin thêm về các trang web lân cận có thể giúp dự đoán của chúng tôi tốt hơn.

4. Các mô hình Máy Học Tập trong Công nghệ Thông tin Địa học

4.1 Mô hình ***Empirical Bayesian Kriging*** (EBK)

Empirical Bayesian Kriging (EBK) là một phương pháp nội suy thống kê địa lý tự động hóa các khía cạnh khó nhất của việc xây dựng một mô hình kriging hợp lệ. Các phương pháp kriging khác trong Geostatistical Analyst yêu cầu bạn điều chỉnh các thông số theo cách thủ công để nhận được kết quả chính xác, nhưng EBK sẽ tự động tính toán các thông số này thông qua quá trình tập hợp con và mô phỏng.



Hình 14: *Empirical Bayesian Kriging* (EBK)

Empirical Bayesian kriging cũng khác với các phương pháp kriging khác bằng cách tính toán sai số được đưa ra bằng cách ước lượng bán biểu đồ cơ bản. Các phương pháp kriging khác tính toán bán biểu đồ từ các vị trí dữ liệu đã biết và sử dụng biểu đồ bán biến đơn này để đưa ra dự đoán tại các vị trí chưa biết; quá trình này mặc nhiên giả định rằng biểu đồ bán biến ước lượng là biểu đồ bán biến thực sự cho vùng nội suy. Bằng cách không tính đến độ không đảm bảo của ước lượng bán biểu đồ, các phương pháp kriging khác đánh giá thấp các sai số chuẩn của dự đoán.

Empirical Bayesian kriging có một số ưu điểm và nhược điểm so với các phương pháp nội suy khác.

Ưu điểm:

- Yêu cầu mô hình tương tác tối thiểu.
- Sai số chuẩn của dự đoán chính xác hơn các phương pháp kriging khác.
- Cho phép dự đoán chính xác dữ liệu không ổn định vừa phải.
- Chính xác hơn các phương pháp kriging khác cho các tập dữ liệu nhỏ.

Khuyết điểm:

- Thời gian xử lý tăng nhanh khi số lượng điểm đầu vào, kích thước tập hợp con hoặc hệ số chồng chéo tăng lên. Việc áp dụng một phép biến đổi cũng sẽ làm tăng thời gian xử lý, đặc biệt nếu K-Bessel hoặc K-Bessel Detrended được chọn cho loại mô hình bán biểu đồ. Các thông số này được mô tả trong các phần tiếp theo của chủ đề này.
- Quá trình xử lý chậm hơn so với các phương pháp kriging khác, đặc biệt là khi xuất ra raster.
- Không có hiệu chỉnh động cơ và dị hướng.
- Phép biến đổi Log Empirical đặc biệt nhạy cảm với các ngoại lệ. Nếu bạn sử dụng phép biến đổi này với dữ liệu có chứa các giá trị ngoại lệ, bạn có thể nhận được các dự đoán là các đơn hàng có độ lớn lớn hơn hoặc nhỏ hơn giá trị của các điểm đầu vào của bạn. Tham số này được mô tả trong phần Biến đổi bên dưới.

4.2 Mô hình ***Support Vector Machine*** (SVM)

Phân loại dữ liệu là một nhiệm vụ phổ biến trong học máy. Giả sử một số điểm dữ liệu đã cho, mỗi điểm thuộc một trong hai lớp và mục tiêu là quyết định lớp nào mà một điểm dữ liệu mới sẽ nằm trong. Trong trường hợp là máy vectơ hỗ trợ, một điểm dữ liệu được xem như là một vectơ ρ-chiều và chúng tôi muốn biết liệu chúng tôi có thể tách các điểm như vậy bằng siêu phẳng có $\rho - 1$ chiều hay không. Đây được gọi là bộ phân loại tuyến tính. Có nhiều siêu máy bay có thể phân loại dữ liệu. Một sự lựa chọn hợp lý vì siêu phẳng tốt nhất là siêu phẳng đại diện cho khoảng cách lớn nhất, hay lề, giữa hai lớp. Vì vậy, chúng tôi chọn siêu phẳng sao cho khoảng cách từ nó đến điểm dữ liệu gần nhất ở mỗi bên là tối đa. Nếu một siêu phẳng như vậy tồn tại, nó được gọi là siêu phẳng có lề tối đa và bộ phân loại tuyến tính mà nó định nghĩa được gọi là bộ phân loại lề tối đa; hoặc tương đương, perceptron của độ ổn định tối ưu.

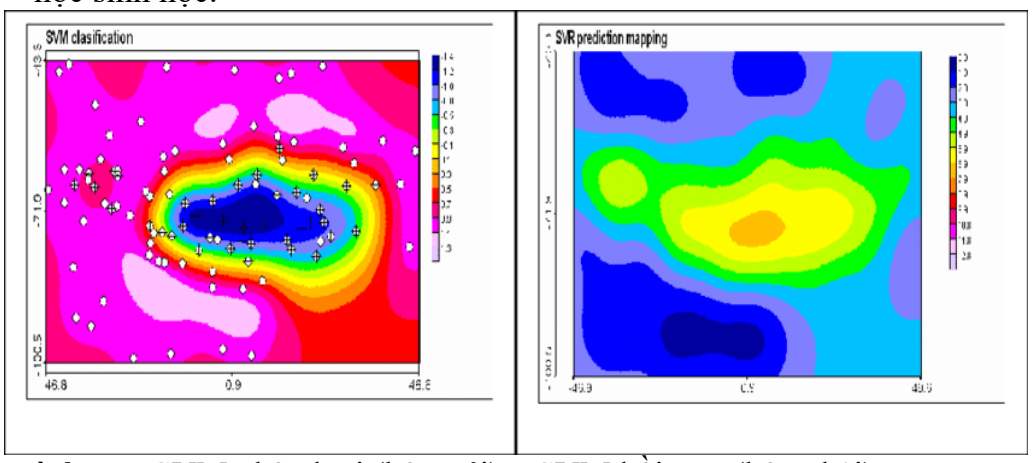
Chính thức hơn, máy vector hỗ trợ tạo ra một siêu phẳng hoặc tập hợp các siêu phẳng trong một không gian chiều cao hoặc vô hạn, có thể được sử dụng để phân loại, hồi quy hoặc các nhiệm vụ khác như phát hiện các giá trị ngoại lai. Theo trực quan, một siêu phẳng có khoảng cách lớn nhất đến điểm dữ liệu huấn luyện gần nhất của bất kỳ lớp nào (còn gọi là lề chức năng), vì nói chung, lề càng lớn thì lỗi tổng quát của bộ phân loại càng thấp.

Trong khi vấn đề ban đầu có thể được phát biểu trong một không gian hữu hạn chiều, nó thường xảy ra rằng các tập phân biệt không thể phân tách tuyến tính trong không gian đó. Vì lý do này, người ta đề xuất rằng không gian hữu hạn chiều ban đầu được ánh xạ thành một không gian có chiều cao hơn nhiều, có lẽ là làm cho việc phân tách trong không gian đó trở nên dễ dàng hơn. Để giữ cho tải tính toán hợp lý, các ánh xạ được sử dụng bởi lược đồ SVM được thiết kế để đảm bảo rằng các sản phẩm dấu chấm của các cặp vectơ dữ liệu đầu vào có thể được tính toán dễ dàng theo các biến trong không gian gốc, bằng cách xác định chúng theo hàm nhân $k(x, y)$ được chọn cho phù hợp với vấn đề. Các siêu phẳng trong không gian chiều cao hơn được định nghĩa là tập hợp các điểm có tích điểm với một vectơ trong không gian đó là hằng số, trong đó tập các vectơ như vậy là một tập các vectơ trực giao (và do đó nhỏ nhất) xác định một siêu phẳng. Các vectơ xác định các siêu mặt phẳng có thể được chọn để kết hợp tuyến tính với các tham số α_i của hình ảnh các vectơ đặc trưng x_i xuất hiện trong kho dữ liệu. Với sự lựa chọn siêu phẳng này, các điểm x trong không gian đối tượng được ánh xạ vào siêu phẳng được xác định bởi quan hệ $\sum_i \alpha_i k(x_i, x) = const$ Lưu ý rằng nếu $k(x, y)$ trở nên nhỏ hơn khi y phát triển xa hơn so với x, mỗi số hạng tính bằng tổng đo mức độ gần của điểm kiểm tra x với điểm cơ sở dữ liệu tương ứng x_i. Bằng cách này, tổng các hạt nhân ở trên có thể được sử dụng để đo mức độ gần tương đối của mỗi điểm kiểm tra với các điểm dữ liệu bắt nguồn từ một hoặc điểm khác của các tập hợp được phân biệt. Lưu ý rằng tập hợp các điểm x được ánh xạ vào bất kỳ siêu phẳng nào do đó có thể khá phức tạp, cho phép phân biệt phức tạp hơn nhiều giữa các tập hợp không lồi chút nào trong không gian ban đầu.

SVM có thể được sử dụng để giải quyết các vấn đề khác nhau trong thế giới thực:

- SVM rất hữu ích trong việc phân loại văn bản và siêu văn bản, vì ứng dụng của chúng có thể làm giảm đáng kể nhu cầu về các phiên bản đào tạo được gắn nhãn trong cả cài đặt quy nạp và chuyển đổi tiêu chuẩn. Một số phương pháp để phân tích ngữ nghĩa nông dựa trên máy vectơ hỗ trợ.
- Việc phân loại hình ảnh cũng có thể được thực hiện bằng SVM. Kết quả thử nghiệm cho thấy SVM đạt được độ chính xác tìm kiếm cao hơn đáng kể so với các lược đồ sàng lọc truy vấn truyền thống chỉ sau ba đến bốn vòng phản hồi về mức độ liên quan. Điều này cũng đúng đối với các hệ thống phân đoạn hình ảnh, bao gồm cả những hệ thống sử dụng SVM phiên bản sửa đổi sử dụng cách tiếp cận đặc quyền theo đề xuất của Vapnik.
- Phân loại dữ liệu vệ tinh như dữ liệu SAR sử dụng SVM có giám sát.
- Các ký tự viết tay có thể được nhận dạng bằng SVM.

- Thuật toán SVM đã được áp dụng rộng rãi trong sinh học và các ngành khoa học khác. Chúng đã được sử dụng để phân loại protein với tới 90% các hợp chất được phân loại chính xác. Kiểm tra hoán vị dựa trên trọng số SVM đã được đề xuất như một cơ chế để giải thích các mô hình SVM. Trọng số máy vectơ hỗ trợ cũng đã được sử dụng để giải thích các mô hình SVM trong quá khứ. Giải thích hậu kỳ của mô hình máy vector hỗ trợ nhằm xác định các đặc điểm được mô hình sử dụng để đưa ra dự đoán là một lĩnh vực nghiên cứu tương đối mới có ý nghĩa đặc biệt trong khoa học sinh học.



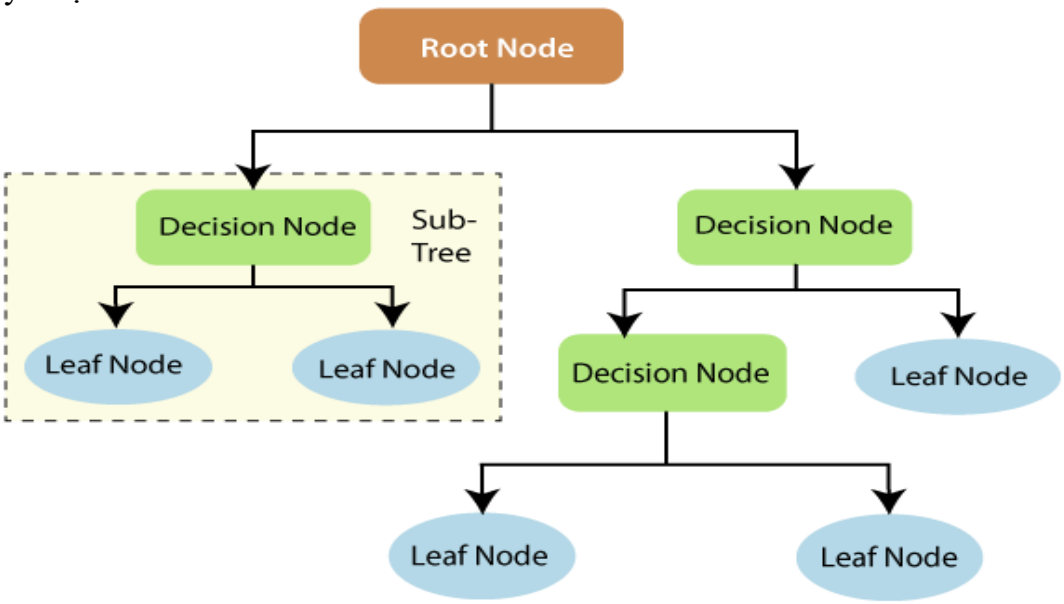
Hình 15: SVM phân loại (bên trái) – SVM hồi quy (bên phải)

4.3 Mô hình ***Random Forest Classification*** (RFC)

Khu rừng ngẫu nhiên là một kỹ thuật máy học được sử dụng để giải quyết các vấn đề hồi quy và phân loại. Nó sử dụng phương pháp học tập hợp, là một kỹ thuật kết hợp nhiều bộ phân loại để đưa ra giải pháp cho các vấn đề phức tạp. Một thuật toán rừng ngẫu nhiên bao gồm nhiều cây quyết định. 'Rừng' được tạo bởi thuật toán rừng ngẫu nhiên được đào tạo thông qua đóng gói hoặc tổng hợp bootstrap. Bagging là một thuật toán meta tổng hợp giúp cải thiện độ chính xác của các thuật toán học máy. Thuật toán thiết lập kết quả dựa trên các dự đoán của cây quyết định. Nó dự đoán bằng cách lấy giá trị trung bình hoặc giá trị trung bình của sản lượng từ các cây khác nhau. Tăng số lượng cây sẽ tăng độ chính xác của kết quả. Một khu rừng ngẫu nhiên loại bỏ những hạn chế của thuật toán cây quyết định. Nó làm giảm việc trang bị quá nhiều bộ dữ liệu và tăng độ chính xác. Nó tạo ra các dự đoán mà không yêu cầu nhiều cấu hình trong các gói

Hoạt động của thuật toán rừng ngẫu nhiên: Cây quyết định là các khối xây dựng của một thuật toán rừng ngẫu nhiên. Cây quyết định là một kỹ thuật hỗ trợ quyết định tạo thành một cấu trúc giống như cây. Tổng quan về cây quyết định sẽ giúp chúng ta hiểu cách hoạt động của các thuật toán rừng ngẫu nhiên. Cây quyết định bao gồm ba thành phần: nút quyết định, nút lá và nút gốc. Thuật toán cây quyết định chia tập dữ liệu huấn luyện thành các nhánh, tập dữ liệu này sẽ tách biệt thành các nhánh khác. Trình tự này tiếp tục cho đến khi đạt được một nút lá. Nút lá không thể được phân tách thêm. Các nút trong cây quyết định đại diện cho các thuộc tính được sử dụng để dự đoán kết quả. Các nút quyết định

cung cấp một liên kết đến các lá. Sơ đồ sau đây cho thấy ba loại nút trong cây quyết định.

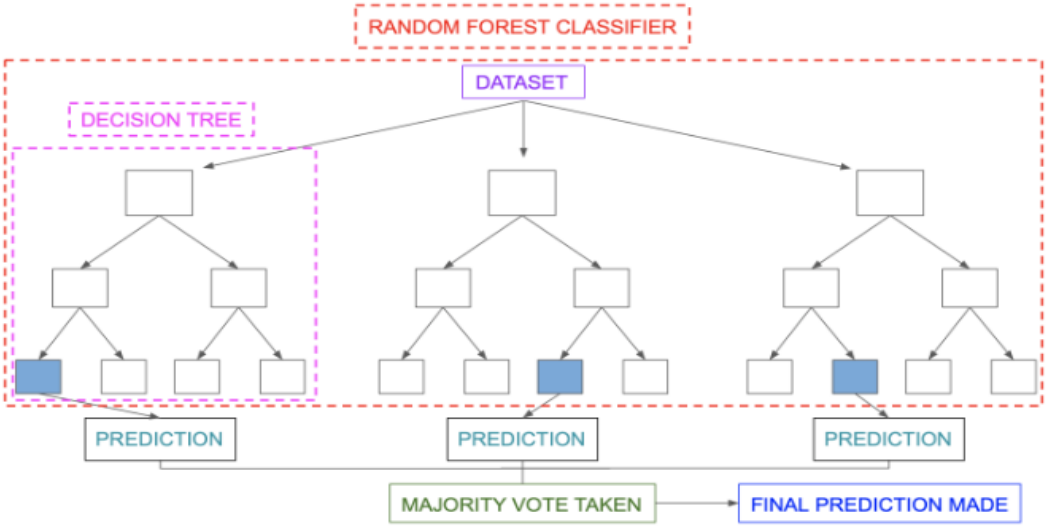


Hình 16: Cấu trúc cây quyết định

Lý thuyết thông tin có thể cung cấp thêm thông tin về cách thức hoạt động của cây quyết định. Entropy và thông tin thu được là các khối xây dựng của cây quyết định. Tổng quan về các khái niệm cơ bản này sẽ nâng cao hiểu biết của chúng ta về cách xây dựng cây quyết định. Entropy là một số liệu để tính toán độ không đảm bảo. Mức tăng thông tin là thước đo mức độ không chắc chắn trong biến mục tiêu được giảm xuống như thế nào, với một tập hợp các biến độc lập. Khái niệm thu thập thông tin liên quan đến việc sử dụng các biến độc lập (đặc trưng) để thu được thông tin về một biến mục tiêu (lớp). Entropy của biến mục tiêu (Y) và entropy có điều kiện của Y (cho trước X) được sử dụng để ước tính mức tăng thông tin. Trong trường hợp này, entropy có điều kiện bị trừ khỏi entropy của Y. Mức tăng thông tin được sử dụng trong việc huấn luyện cây quyết định. Nó giúp giảm sự không chắc chắn ở những cây này. Mức tăng thông tin cao có nghĩa là mức độ không chắc chắn cao (entropy thông tin) đã bị loại bỏ. Entropy và thu thập thông tin rất quan trọng trong việc tách các nhánh, đây là một hoạt động quan trọng trong việc xây dựng cây quyết định. Sự khác biệt chính giữa thuật toán cây quyết định và thuật toán rừng ngẫu nhiên là việc thiết lập các nút gốc và các nút tách biệt được thực hiện ngẫu nhiên trong thuật toán sau. Khu rừng ngẫu nhiên sử dụng phương pháp đóng bao để tạo ra dự đoán cần thiết. Việc đóng gói bao gồm việc sử dụng các mẫu dữ liệu khác nhau (dữ liệu đào tạo) thay vì chỉ một mẫu. Tập dữ liệu đào tạo bao gồm các quan sát và tính năng được sử dụng để đưa ra dự đoán. Các cây quyết định tạo ra các đầu ra khác nhau, tùy thuộc vào dữ liệu huấn luyện được cung cấp cho thuật toán rừng ngẫu nhiên. Các kết quả đầu ra này sẽ được xếp hạng, và kết quả cao nhất sẽ được chọn làm đầu ra cuối cùng. Thay vì có một cây quyết định duy nhất, khu rừng ngẫu nhiên sẽ có nhiều cây quyết định. Giả sử chúng ta chỉ có bốn cây quyết định. Trong trường hợp này, dữ liệu đào tạo bao gồm các quan sát và tính năng của điện thoại sẽ được chia thành bốn nút gốc. Các nút gốc có thể đại diện cho bốn tính năng có

thể ảnh hưởng đến sự lựa chọn của khách hàng (giá cả, bộ nhớ trong, máy ảnh và RAM). Khu rừng ngẫu nhiên sẽ chia các nút bằng cách chọn các tính năng một cách ngẫu nhiên. Dự đoán cuối cùng sẽ được chọn dựa trên kết quả của 4 cây. Kết quả được chọn bởi hầu hết các cây quyết định sẽ là lựa chọn cuối cùng.

Phân loại với thuật toán rừng ngẫu nhiên: Việc phân loại trong các khu rừng ngẫu nhiên sử dụng một phương pháp luận tổng hợp để đạt được kết quả. Dữ liệu đào tạo được cung cấp để đào tạo các cây quyết định khác nhau. Tập dữ liệu này bao gồm các quan sát và các tính năng sẽ được chọn ngẫu nhiên trong quá trình tách các nút. Hệ thống rừng mưa phụ thuộc vào nhiều loại cây quyết định khác nhau. Mọi cây quyết định đều bao gồm các nút quyết định, nút lá và nút gốc. Nút lá của mỗi cây là đầu ra cuối cùng do cây quyết định cụ thể đó tạo ra. Việc lựa chọn đầu ra cuối cùng tuân theo hệ thống bỏ phiếu đa số. Trong trường hợp này, sản lượng được chọn bởi phần lớn các cây quyết định sẽ trở thành sản lượng cuối cùng của hệ thống rừng mưa. Biểu đồ dưới đây cho thấy một bộ phân loại rừng ngẫu nhiên đơn giản.

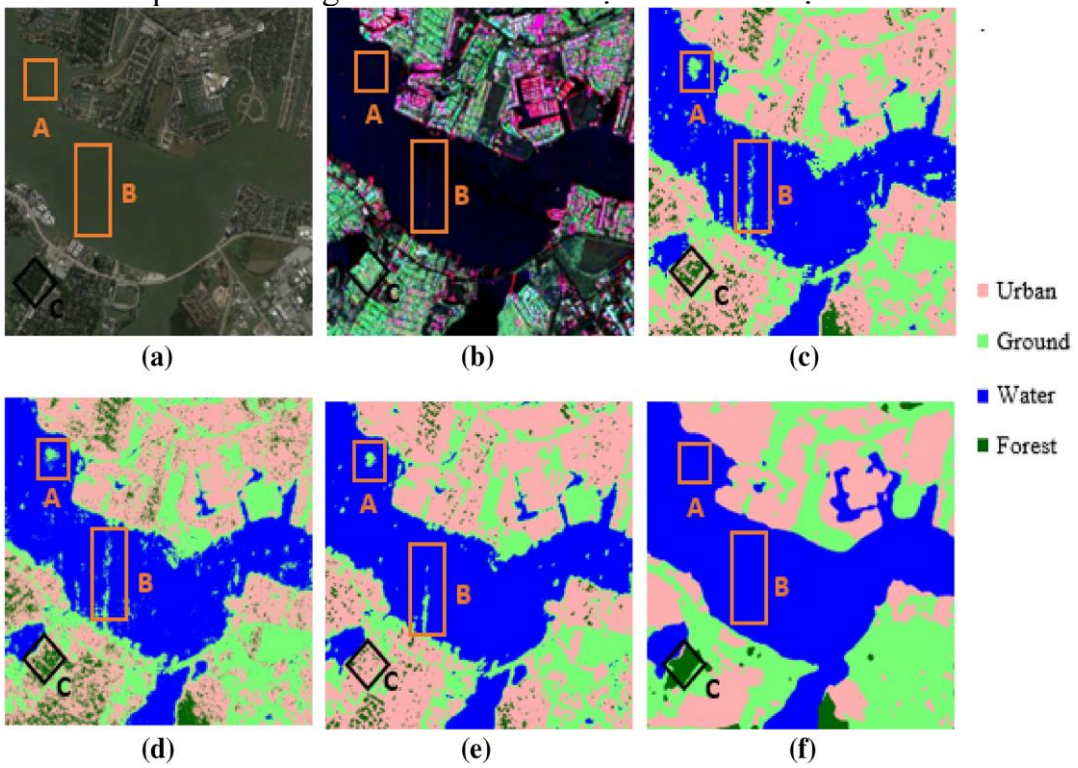


Hình 17: Cấu trúc thuật toán rừng ngẫu nhiên

Hồi quy với thuật toán rừng ngẫu nhiên: Hồi quy là nhiệm vụ khác được thực hiện bởi một thuật toán rừng ngẫu nhiên. Hồi quy rừng ngẫu nhiên tuân theo khái niệm hồi quy đơn giản. Giá trị của các biến phụ thuộc (đặc trưng) và biến độc lập được chuyển trong mô hình rừng ngẫu nhiên. Chúng tôi có thể chạy hồi quy rừng ngẫu nhiên trong các chương trình khác nhau như SAS, R và python. Trong hồi quy rừng ngẫu nhiên, mỗi cây đưa ra một dự đoán cụ thể. Dự đoán trung bình của các cây riêng lẻ là đầu ra của hồi quy. Điều này trái với phân loại rừng ngẫu nhiên, mà sản lượng của nó được xác định theo phương thức của loại cây quyết định. Mặc dù hồi quy rừng ngẫu nhiên và hồi quy tuyến tính theo cùng một khái niệm, chúng khác nhau về chức năng. Hàm của hồi quy tuyến tính là $y = bx + c$, trong đó y là biến phụ thuộc, x là biến độc lập, b là tham số ước lượng và c là hằng số. Chức năng của một hồi quy rừng ngẫu nhiên phức tạp giống như một hộp đen.

4.4 Mô hình ***Image Segmentaion***

Trong xử lý ảnh kỹ thuật số và thị giác máy tính, phân đoạn ảnh là quá trình phân vùng ảnh kỹ thuật số thành nhiều đoạn ảnh, còn được gọi là vùng ảnh hoặc đối tượng ảnh (tập hợp các pixel). Mục tiêu của phân đoạn là đơn giản hóa và / hoặc thay đổi hình ảnh đại diện thành một thứ có ý nghĩa hơn và dễ phân tích hơn. Phân đoạn hình ảnh thường được sử dụng để xác định vị trí các đối tượng và ranh giới (đường thẳng, đường cong, v.v.) trong hình ảnh. Chính xác hơn, phân đoạn hình ảnh là quá trình gán nhãn cho mọi pixel trong một hình ảnh sao cho các pixel có cùng nhãn chia sẻ các đặc điểm nhất định.



Hình 18: Phân đoạn ngữ nghĩa trên dữ liệu SAR

Kết quả của phân đoạn hình ảnh là một tập hợp các phân đoạn bao phủ chung toàn bộ hình ảnh, hoặc một tập hợp các đường bao được trích xuất từ hình ảnh (xem phần phát hiện cạnh). Mỗi pixel trong một vùng tương tự nhau về một số đặc tính hoặc thuộc tính được tính toán, chẳng hạn như màu sắc, cường độ hoặc kết cấu. Các vùng lân cận có màu sắc khác nhau đáng kể so với (các) đặc tính giống nhau. Khi được áp dụng cho một chồng ảnh, điển hình trong hình ảnh y tế, các đường viền thu được sau khi phân đoạn hình ảnh có thể được sử dụng để tạo ra các bản tái tạo 3D với sự trợ giúp của các thuật toán nội suy như các hình khối diễu hành.

Các nhóm thuộc mô hình phân đoạn ảnh gồm:

- Phân đoạn ngữ nghĩa (semantic segmentation) là một cách tiếp cận phát hiện, đối với mỗi pixel, lớp thuộc về đối tượng. Ví dụ: khi tất cả mọi người trong một hình được phân đoạn thành một đối tượng và nền là một đối tượng.
- Phân đoạn đối tượng (instance segmentation) là một phương pháp xác định, đối với mỗi pixel, một thể hiện thuộc về đối tượng. Nó phát hiện

từng đối tượng quan tâm khác nhau trong hình ảnh. Ví dụ, khi mỗi người trong một hình được phân đoạn thành một đối tượng riêng lẻ.

- Phân đoạn khái quát (panotic segmentation) kết hợp phân đoạn ngữ nghĩa và phân đoạn thể hiện. Giống như phân đoạn theo ngữ nghĩa, phân đoạn toàn cảnh là một cách tiếp cận xác định, đối với mỗi pixel, lớp thuộc về. Không giống như phân đoạn ngữ nghĩa, phân đoạn toàn cảnh phân biệt các trường hợp khác nhau của cùng một lớp.

Để thực hiện phân đoạn ảnh, các mô hình thường sử dụng một trong các phương pháp phân đoạn sau:

Phân đoạn theo ngưỡng (thresholding): Phương pháp phân đoạn ảnh đơn giản nhất được gọi là phương pháp phân ngưỡng. Phương pháp này dựa trên mức clip (hoặc giá trị ngưỡng) để biến hình ảnh thang xám thành hình ảnh nhị phân. Chìa khóa của phương pháp này là chọn giá trị ngưỡng (hoặc các giá trị khi nhiều cấp được chọn). Một số phương pháp phổ biến được sử dụng trong công nghiệp bao gồm phương pháp entropy cực đại, ngưỡng biểu đồ cân bằng, phương pháp Otsu (phương sai tối đa) và phân cụm k-mean.

Phân đoạn theo histogram: Các phương pháp dựa trên biểu đồ rất hiệu quả so với các phương pháp phân đoạn hình ảnh khác vì chúng thường chỉ yêu cầu một lần đi qua các pixel. Trong kỹ thuật này, biểu đồ được tính toán từ tất cả các pixel trong hình ảnh và các đỉnh và vùng lõm trong biểu đồ được sử dụng để định vị các cụm trong hình ảnh. Màu sắc hoặc cường độ có thể được sử dụng làm thước đo. Một cải tiến của kỹ thuật này là áp dụng đệ quy phương pháp tìm kiếm biểu đồ cho các cụm trong hình ảnh để chia chúng thành các cụm nhỏ hơn. Thao tác này được lặp lại với các cụm nhỏ hơn và nhỏ hơn cho đến khi không còn cụm nào được hình thành. Một nhược điểm của phương pháp tìm kiếm biểu đồ là có thể khó xác định các đỉnh và thung lũng quan trọng trong hình ảnh. Các phương pháp tiếp cận dựa trên biểu đồ cũng có thể nhanh chóng được điều chỉnh để áp dụng cho nhiều khung hình, trong khi vẫn duy trì hiệu quả chuyển đơn của chúng. Biểu đồ có thể được thực hiện theo nhiều kiểu khi nhiều khung được xem xét. Phương pháp tương tự được thực hiện với một khung có thể được áp dụng cho nhiều khung và sau khi kết quả được hợp nhất, các đỉnh và thung lũng trước đây khó xác định có nhiều khả năng phân biệt hơn. Biểu đồ cũng có thể được áp dụng trên cơ sở mỗi pixel trong đó thông tin kết quả được sử dụng để xác định màu thường xuyên nhất cho vị trí pixel. Cách tiếp cận này phân đoạn dựa trên các đối tượng đang hoạt động và môi trường tĩnh, dẫn đến một loại phân đoạn khác hữu ích trong việc theo dõi video.

Phân đoạn theo đường biên (edge detection): Phát hiện cạnh là một lĩnh vực được phát triển tốt trong quá trình xử lý hình ảnh. Ranh giới và các cạnh của vùng có liên quan chặt chẽ với nhau, vì thường có sự điều chỉnh mạnh về cường độ tại các ranh giới vùng. Do đó, các kỹ thuật phát hiện cạnh đã được sử dụng làm cơ sở của một kỹ thuật phân đoạn khác. Các cạnh được xác định bằng phát hiện cạnh thường bị ngắt kết nối. Tuy nhiên, để phân đoạn một đối tượng từ một hình ảnh, một đối tượng cần có ranh giới vùng kín. Các cạnh mong muốn là ranh giới giữa các đối tượng như vậy hoặc các đơn vị phân loại không gian. Các đơn

vị phân loại không gian là các hạt thông tin, bao gồm một vùng pixel sắc nét, được đặt ở các mức trừu tượng trong một kiến trúc cảnh lồng nhau có phân cấp. Chúng tương tự như ký hiệu tâm lý học Gestalt của hình nền, nhưng được mở rộng để bao gồm tiền cảnh, các nhóm đối tượng, đối tượng và các phần đối tượng nổi bật. Các phương pháp phát hiện cạnh có thể được áp dụng cho vùng không gian-đơn vị phân loại, giống như cách chúng sẽ được áp dụng cho một hình bóng. Phương pháp này đặc biệt hữu ích khi cạnh bị ngắt kết nối là một phần của đường bao ảo. Phương pháp phân đoạn cũng có thể được áp dụng cho các cạnh thu được từ máy dò cạnh. Lindeberg và Li đã phát triển một phương pháp tích hợp phân đoạn các cạnh thành các đoạn cạnh thẳng và cong để nhận dạng đối tượng dựa trên các bộ phận, dựa trên tiêu chí độ dài mô tả tối thiểu (MDL) được tối ưu hóa bằng phương pháp giống như tách và hợp nhất với các điểm ngắt ứng viên thu được từ các dấu hiệu đường giao nhau bổ sung để có được nhiều điểm có khả năng hơn để xem xét các phân vùng thành các phân đoạn khác nhau.

Region-growing methods: Phương pháp tăng trưởng theo vùng chủ yếu dựa vào giả định rằng các pixel lân cận trong một vùng có giá trị tương tự. Quy trình phổ biến là so sánh một pixel với các pixel lân cận của nó. Nếu một tiêu chí tương tự được đáp ứng, pixel có thể được đặt để thuộc cùng một cụm với một hoặc nhiều điểm ảnh lân cận. Việc lựa chọn tiêu chí tương tự là rất quan trọng và kết quả bị ảnh hưởng bởi nhiễu trong mọi trường hợp. Phương pháp Hợp nhất Vùng Thống kê (SRM) bắt đầu bằng cách xây dựng biểu đồ các pixel sử dụng tính liên kết 4 với các cạnh được tính trọng số bằng giá trị tuyệt đối của chênh lệch cường độ. Ban đầu mỗi pixel tạo thành một vùng pixel duy nhất. Sau đó SRM sắp xếp các cạnh đó trong một hàng đợi ưu tiên và quyết định có hợp nhất các vùng hiện tại thuộc về các pixel cạnh hay không bằng cách sử dụng một vị từ thống kê. Một phương pháp trồng theo vùng là phương pháp trồng vùng bằng hạt. Phương pháp này lấy một tập hợp các hạt giống làm đầu vào cùng với hình ảnh. Các hạt đánh dấu từng đối tượng được phân đoạn. Các vùng được phát triển lặp đi lặp lại bằng cách so sánh tất cả các pixel lân cận chưa được phân bổ với các vùng. Sự khác biệt giữa giá trị cường độ của pixel và giá trị trung bình của vùng, δ, được sử dụng làm thước đo mức độ tương đồng. Pixel có sự khác biệt nhỏ nhất được đo theo cách này được gán cho vùng tương ứng. Quá trình này tiếp tục cho đến khi tất cả các pixel được gán cho một vùng. Bởi vì việc trồng vùng gieo hạt yêu cầu hạt giống như đầu vào bổ sung, kết quả phân đoạn phụ thuộc vào việc lựa chọn hạt giống và nhiễu trong hình ảnh có thể khiến hạt giống được đặt kém. Một phương pháp trồng trọt theo vùng khác là phương pháp trồng vùng kín. Nó là một thuật toán được sửa đổi không yêu cầu các hạt giống rõ ràng. Nó bắt đầu với một vùng A_1 — pixel được chọn ở đây không ảnh hưởng rõ rệt đến phân đoạn cuối cùng. Ở mỗi lần lặp lại, nó sẽ xem xét các pixel lân cận theo cách giống như vùng được gieo hạt đang phát triển. Nó khác với vùng gieo hạt đang phát triển ở chỗ nếu δ tối thiểu nhỏ hơn ngưỡng xác định trước T thì vùng đó sẽ được thêm vào vùng tương ứng A_j. Nếu không, thì pixel được coi là khác với tất cả các vùng hiện tại A_i và một vùng mới A_{n+1} được tạo bằng pixel này . Một biến thể của kỹ thuật này, được đề xuất bởi Haralick và Shapiro (1985), dựa trên cường độ pixel. Giá trị trung bình và độ phân tán của vùng cũng như cường

độ của pixel ứng viên được sử dụng để tính toán thống kê thử nghiệm. Nếu thống kê thử nghiệm đủ nhỏ, pixel sẽ được thêm vào vùng, giá trị trung bình và độ phân tán của vùng sẽ được tính toán lại. Nếu không, pixel sẽ bị từ chối và được sử dụng để tạo một vùng mới. Một phương pháp phát triển vùng đặc biệt được gọi là γ-connected segmentation (xem thêm lambda-connectness). Nó dựa trên cường độ pixel và đường dẫn liên kết vùng lân cận. Mức độ kết nối (kết nối) được tính toán dựa trên một đường dẫn được hình thành bởi các pixel. Đối với một giá trị nhất định của γ, hai pixel được gọi là γ-connected nếu có một đường dẫn liên kết hai pixel đó và độ kết nối của đường dẫn này ít nhất là γ. γ-connectedness là một quan hệ tương đương. Phân đoạn tách và hợp nhất dựa trên phân vùng quadtree của một hình ảnh. Nó đôi khi được gọi là phân đoạn quadtree. Phương thức này bắt đầu từ gốc của cây đại diện cho toàn bộ hình ảnh. Nếu nó được tìm thấy là không đồng nhất (không đồng nhất), thì nó được tách thành bốn hình vuông con (quá trình tách), v.v. Ngược lại, nếu bốn hình vuông con là đồng nhất, chúng được hợp nhất như một số thành phần được kết nối (quá trình hợp nhất). Nút trong cây là một nút được phân đoạn. Quá trình này tiếp tục đệ quy cho đến khi không thể tách hoặc hợp nhất thêm nữa. Khi cấu trúc dữ liệu đặc biệt tham gia vào việc triển khai thuật toán của phương pháp, độ phức tạp về thời gian của nó có thể đạt đến $O(nlog(n))$, một thuật toán tối ưu của phương pháp.

Tài liệu tham khảo

[1]. Marwick, Ben; Hiscock, Peter; Sullivan, Marjorie; Hughes, Philip (July 2017). "Landform boundary effects on Holocene forager landscape use in arid South Australia". Journal of Archaeological Science: Reports. 19: 864–874. doi:10.1016/j.jasrep.2017.07.004. S2CID 134572456.

[2]. Longley, Paul A.; Goodchilde, Michael F.; Maguire, David J.; Rhind, David W. (2015). Geographic Information Systems & Science (4th ed.). Wiley.

[3]. C. Bayindir; J. D. Frost; C. F. Barnes (January 2018). "Assessment and enhancement of SAR noncoherent change detection of sea-surface oil spills". IEEE J. Ocean. Eng. 43 (1): 211–220.Bibcode:2018IJOE...43..211B. doi:10.1109/JOE.2017.2714818. S2CID 44706251.

[4]. Makki, Ihab; Younes, Rafic; Francis, Clovis; Bianchi, Tiziano; Zucchetti, Massimo (1 February 2017). "A survey of landmine detection using hyperspectral imaging". ISPRS Journal of Photogrammetry and Remote Sensing. 124: 40–53.

[5]. Ditter, R., Haspel, M., Jahn, M., Kollar, I., Siegmund, A., Viehrig, K., Volz, D., Siegmund, A. (2012) Geospatial technologies in school – theoretical concept and practical implementation in K-12 schools. In: International Journal of Data Mining, Modelling and Management (IJDMMM): FutureGIS: Riding the Wave of a Growing Geospatial Technology Literate Society; Vol. X

[6]. Cobb, Peter J.; Earley-Spadoni, Tiffany; Dames, Philip (2019). "Centimeter-Level Recording for All: Field Experimentation with New, Affordable Geolocation Technology". Advances in Archaeological Practice. 7 (4): 353–365.

[7]. K. Iwasaki, K. Yamazawa, and N. Yokoya. An indexing system for photos based on shooting position and orientation with geographic database. In IEEE International Conference on Multimedia and Expo, ICME 2005, pages 390–393, 2005. (doi:10.1109/ICME.2005.1521442)

[8]. Russell, Stuart J.; Norvig, Peter (2003), Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2.

[9]. Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012). Foundations of Machine Learning. The MIT Press. ISBN 9780262018258.

[10]. Mitchell, T. (1997). Machine Learning. McGraw Hill. p. 2. ISBN 978-0-07-042807-2.

[11]. Hinton, G.E. (2009). "Deep belief networks". Scholarpedia. 4 (5): 5947.

[12]. Hassoun, Mohamad H. (1995). Fundamentals of Artificial Neural Networks. MIT Press. p. 48. ISBN 978-0-262-08239-6.

[13]. Socher, Richard; Manning, Christopher. "Deep Learning for NLP" (PDF). Archived (PDF) from the original on 6 July 2014. Retrieved 26 October 2014.

[14]. Raina, Rajat; Madhavan, Anand; Ng, Andrew Y. (2009). "Large-scale Deep Unsupervised Learning Using Graphics Processors". Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09. New York, NY, USA: ACM: 873–880. CiteSeerX 10.1.1.154.372. doi:10.1145/1553374.1553486. ISBN 9781605585161. S2CID 392458.

[15]. Sze, Vivienne; Chen, Yu-Hsin; Yang, Tien-Ju; Emer, Joel (2017). "Efficient Processing of Deep Neural Networks: A Tutorial and Survey". arXiv:1703.09039

[16]. Ciresan, D. C.; Meier, U.; Masci, J.; Gambardella, L. M.; Schmidhuber, J. (2011). "Flexible, High Performance Convolutional Neural Networks for Image Classification" (PDF). International Joint Conference on Artificial Intelligence. doi:10.5591/978-1-57735-516-8/ijcai11-210

[17]. Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng. "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations" Proceedings of the 26th Annual International Conference on Machine Learning, 2009.

[18]. Y. Bengio; A. Courville; P. Vincent (2013). "Representation Learning: A Review and New Perspectives". IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (8): 1798–1828. arXiv:1206.5538. doi:10.1109/tpami.2013.50. PMID 23787338. S2CID 393948.

[19]. Zimek, Arthur; Schubert, Erich (2017), "Outlier Detection", Encyclopedia of Database Systems, Springer New York, pp. 1–5, doi:10.1007/978-1-4899-7993-3_80719-1,ISBN: 781489979933

[20]. Zhang, Jun; Zhan, Zhi-hui; Lin, Ying; Chen, Ni; Gong, Yue-jiao; Zhong, Jing-hui; Chung, Henry S.H.; Li, Yun; Shi, Yu-hui (2011). "Evolutionary Computation Meets Machine Learning: A Survey". Computational Intelligence Magazine. 6 (4): 68–75. doi:10.1109/mci.2011.942584. S2CID 6760276.