

Trường Nguyễn & Dương Đỗ (2022). Tầm quan trọng của học máy trong phân tích dữ liệu lớn. *Đặc san Nghiên cứu Chính sách và Phát triển*, 1(2022), 147-152.

Đặc san Nghiên cứu
Chính sách
và Phát triển

© Học viện
Chính sách
và Phát triển, 2022
© CSR, 2022

Bài báo khoa học

Tầm quan trọng của học máy trong phân tích dữ liệu lớn

Nguyễn Hữu Xuân Trường (TS.)

Đỗ Thế Dương (ThS.)

Học viện Chính sách và Phát triển.

Tóm tắt. Cuộc cách mạng công nghiệp 4.0 hiện nay đã dẫn tới sự bùng nổ dữ liệu trên toàn thế giới và tạo ra rất nhiều kho dữ liệu lớn (Big Data). Ẩn chứa trong khối dữ liệu lớn này là những thông tin vô cùng có ích nhưng thật không dễ dàng để trích xuất được. Một nhiệm vụ tưởng như bất khả thi khi những tập dữ liệu lớn này không thể xử lý được với những công cụ truyền thống, nhưng thật là may mắn khi chúng ta đã có chìa khóa để giải quyết hiệu quả cho vấn đề này, đó chính là Học máy (Machine Learning) - một lĩnh vực của trí tuệ nhân tạo (AI) đang phát triển mạnh trong giai đoạn hiện nay. Trong bài viết này, chúng tôi sẽ trình bày về các vấn đề cơ bản của Học máy và tầm quan trọng của nó trong phân tích dữ liệu lớn.

Ngày nhận bài:
15 tháng 12, 2021
Bản sửa lần 1:
25 tháng 12, 2021
Ngày duyệt bài:
30 tháng 12, 2021

Mã số: ĐS140122

Từ khóa: học máy, trí tuệ nhân tạo, dữ liệu lớn, phân tích dữ liệu.

Abstract. Big data was produced as a result of the fourth industrial revolution, which resulted in an explosion of data worldwide. There is incredibly relevant information hidden within this extensive data, but extracting it is not easy. A seemingly insurmountable task when enormous data cannot be processed using typical data processing techniques, but fortunately, Machine Learning, a subfield of artificial intelligence that is rapidly evolving, holds the key to solving this challenge efficiently. This paper will go through the fundamentals of machine learning and how it applies to big data analytics.

Keywords: machine learning, artificial intelligence, big data, data analysis.

1. Giới thiệu về Học máy và Dữ liệu lớn

Vào thời điểm năm 2018, ước tính mỗi ngày chúng ta tạo ra khoảng 2,5 tỷ Byte dữ liệu và theo thống kê thì dữ liệu sinh ra trong khoảng hai năm liên tiếp xấp xỉ bằng dữ liệu được sinh từ tất cả các năm trước cộng lại. Thêm vào đó, chi phí cho việc tạo ra và lưu trữ dữ liệu ngày một rẻ hơn, qua thời gian thu thập đã có rất nhiều kho dữ liệu lớn được tạo ra. Dữ liệu lớn có thể hiểu là một tập dữ liệu có kích thước rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được, tuy nhiên nếu trích xuất thành công thì dữ liệu lớn sẽ mang lại nhiều thông tin rất quý giá cho chúng ta (Đặng Thị Hằng, 2019).

Trước đây chúng ta sử dụng dữ liệu chủ yếu mang tính lưu trữ thông tin, hầu như chúng ta đã bỏ quên những tri thức vô cùng có ích được chứa trong các kho dữ liệu lớn. Vậy vấn đề đặt ra là chúng ta phải tìm những tri thức có ích trong đồng dữ liệu khổng lồ đó như thế nào? Tất nhiên là với sức người thì chúng ta không thể mò kim đáy bể trong đồng dữ liệu khổng lồ, nhưng với máy tính vô tri vô giác thì làm sao để biết được tri thức là gì mà tìm? Một nhiệm vụ tưởng như bất khả thi khi những tập dữ liệu lớn này không thể xử lý được với những công cụ truyền thống, nhưng thật là may mắn khi chúng ta đã có chìa khóa để giải quyết hiệu quả cho vấn đề này, đó chính là Học máy - một lĩnh vực của trí tuệ nhân tạo đang phát triển mạnh trong giai đoạn hiện nay.

Học máy được giới thiệu vào cuối những năm 1950, là lĩnh vực nghiên cứu khoa học về các thuật toán và mô hình thống kê mà các hệ thống máy tính sử dụng

để thực hiện một nhiệm vụ cụ thể không cần sử dụng các hướng dẫn rõ ràng, thay vào đó là dựa vào các mẫu và suy luận. Các thuật toán học máy xây dựng một mô hình toán học dựa trên dữ liệu mẫu được gọi là “dữ liệu huấn luyện”, để đưa ra dự đoán hoặc quyết định mà không phải lập trình cụ thể để thực hiện nhiệm vụ (Phạm Ngọc Thọ, 2019).

Có thể lấy ví dụ đơn giản về việc học dựa trên dữ liệu mẫu như sau:

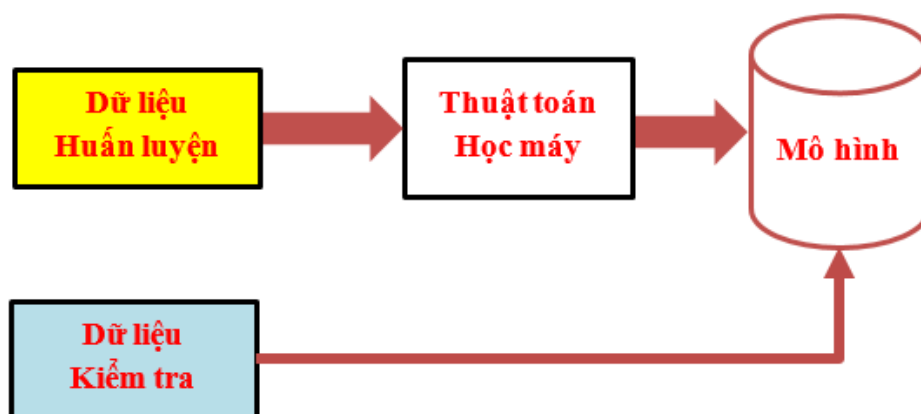
$X = \{2, 3, 5, 6\}$ và tương ứng

$Y = \{3, 5, 9, 11\}$

Câu hỏi đặt ra là với $X = 4$ thì $Y = ?$

Vấn đề trên giống như một câu hỏi về IQ, và với khả năng suy luận của con người thì cũng không quá khó khăn để dự đoán là với $X = 4$ thì $Y = 7$ (suy luận từ dữ liệu thì đoán rằng liên hệ giữa Y và X theo hàm $Y = 2X - 1$). Tuy nhiên để cho máy tính có thể suy luận giống như người thì cần phải sử dụng những thuật toán của Học máy để học bộ dữ liệu mẫu cho trước, sau đó sẽ dự đoán cho dữ liệu mới giống như dự đoán Y với $X = 4$. Dữ liệu mẫu càng nhiều thì việc học càng tốt (kết quả càng chính xác).

Một ví dụ khác nữa cho việc học qua dữ liệu đó là việc nhận dạng chữ viết tay. Giống như một đứa trẻ bình thường tầm 5 - 6 tuổi và chưa biết chữ, để dạy đứa trẻ biết đâu là chữ a thì cũng phải viết mẫu ít nhất vài lần chữ a (và cũng có thể chỉ cho đứa trẻ biết những chữ khác không phải là chữ a như x, y, z) để nó nhận thức được trong đầu mình, sau đó nó hoàn toàn có thể nhận dạng được chữ a trong các văn bản mới, dù là có những chữ a do những người lạ viết mà đứa trẻ chưa được nhìn thấy bao giờ nhưng cơ bản



Hình 1. Quá trình huấn luyện mô hình

là nó vẫn có thể xác định được đây là chữ a. Với máy tính cũng vậy, để có thể nhận dạng được chính xác các chữ a từ nhiều dạng viết khác nhau (qua các ảnh chụp văn bản chẳng hạn) thì chúng ta cũng phải sử dụng rất nhiều mẫu chữ cái khác nhau, sau đó dùng thuật toán của Machine Learning để huấn luyện mô hình có thể nhận biết (dự đoán) đâu là chữ a, đâu không phải là chữ a, đó chính là việc dạy cho máy tính học.

Ngày nay, Học máy nổi lên như một minh chứng của cuộc cách mạng công nghiệp lần thứ tư (1: Động cơ hơi nước; 2: Năng lượng điện; 3: Công nghệ thông tin). Học máy đang len lỏi vào mọi lĩnh vực trong đời sống con người mà có thể chúng ta không nhận ra. Trên thế giới, nhiều tập đoàn công nghệ nổi tiếng đã lấy nền tảng của trí tuệ nhân tạo và học máy cho ra đời nhiều ứng dụng như: Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, hệ thống lọc email spam của Gmail, phát hiện gian lận thẻ tín dụng trong ngân hàng... (Vũ Hữu Tiệp, 2018)

Ưu điểm của Học máy là có thể học tự động từ dữ liệu với tốc độ rất nhanh, độ chính xác cao nên có thể đáp ứng xử lý cho dữ liệu lớn. Học máy và Dữ liệu lớn có mối quan hệ cộng sinh chặt chẽ lẫn nhau, trong khi Dữ liệu lớn giúp cho Học máy có đầy đủ lượng dữ liệu để học và tạo ra các mô hình có độ chính xác cao thì Dữ liệu lớn lại cần Học máy trong việc phân tích xử lý để tìm ra những thông tin hữu ích.

2. Phân loại các thuật toán học máy

2.1. Học có giám sát (Supervised Learning)

Gồm hai giai đoạn: huấn luyện và thử nghiệm. Các bộ dữ liệu được sử dụng trong giai đoạn huấn luyện cần phải có nhãn đã biết. Các thuật toán tìm mối liên quan giữa giá trị đầu vào và nhãn và cố gắng dự đoán các giá trị đầu ra của dữ liệu thử nghiệm. Học có giám sát giải quyết các vấn đề như: Dự báo thời tiết; dự đoán tăng trưởng kinh tế, nhận dạng chữ số, nhận dạng giọng nói, chuẩn đoán bệnh và phát hiện gian lận - lừa đảo... (Vũ Hữu Tiệp, 2018) (Phạm Ngọc Thọ, 2019). Các thuật toán Học có giám sát được chia thành hai loại: phân lớp và hồi quy, mỗi loại cũng gồm những thuật toán khác nhau.

• **Phân lớp (Classification):** phân một đối tượng dữ liệu dựa trên các thuộc tính đã biết vào một lớp có các nhãn cho trước, tập nhãn ở đây là hữu hạn. Chẳng hạn như dựa trên đặc điểm của một người thì ta có thể dự đoán Giới tính (lớp) của họ là Nam hoặc Nữ (tập nhãn có 2 giá trị Nam/Nữ), hoặc có thể dự đoán độ tuổi của họ là thuộc một trong các lớp Thanh niên/ Trung niên/ Cao niên... Một số thuật toán phổ biến của Phân lớp là:

- + Cây quyết định (Decision Tree)
- + K - láng giềng gần nhất (K - Nearest Neighbor)
- + Máy Vector hỗ trợ (Support Vector Machine)
- + Naïve Bayes...

• **Hồi quy (Regression):** dự đoán một giá trị số bằng cách sử dụng các biến đã biết mà không cần phải tìm ra mối quan hệ giữa các biến đó. Điểm khác của Hồi quy so với Phân lớp là tập giá trị dự đoán là rất nhiều và liên tục. Ví dụ của bài toán Hồi quy như dự đoán giá nhà dựa trên các thuộc tính: diện tích, số phòng ngủ, khoảng cách tới trung tâm... Một số thuật toán Hồi quy như hồi quy tuyến tính, hồi quy từng bước, hồi quy đơn biến, hồi quy đa biến...

2.2. Học không giám sát (Unsupervised Learning)

Trái với phương pháp học có giám sát, dữ liệu đầu vào không có nhãn. Các thuật toán trong danh mục này cố gắng xác định các khuôn mẫu trên dữ liệu thử nghiệm và phân cụm dữ liệu hoặc dự đoán các giá trị trong tương lai. Học không giám sát liên quan đến các vấn đề giảm kích thước đặc trưng sử dụng để trực quan hoá dữ liệu

lớn, khai phá đặc trưng, khám phá các cấu trúc ẩn. Hơn nữa, học không giám sát được sử dụng cho các vấn đề phân cụm được áp dụng trong các bài toán như: Hệ thống khuyến nghị; tìm ra phân khúc khách hàng và tiếp thị mục tiêu.

Các thuật toán Học không giám sát cũng được chia thành hai loại là:

• **Phân cụm (Clustering):** K-means, DBSCAN, EM, Hierarchical Clustering...

• **Luật kết hợp (Association Rule)**

2.3. Học bán giám sát (Semi-supervised Learning)

Đây là sự kết hợp của hai phương pháp học có giám sát và không giám sát, trong đó sử dụng cả dữ liệu có sẵn nhãn và dữ liệu không nhãn. Nó hoạt động chủ yếu giống như việc học tập không giám sát với những cải tiến mà một phần dữ liệu được dán nhãn có thể mang lại.

Một ví dụ điển hình của nhóm này là chỉ có một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị) và phần lớn các bức ảnh/văn bản khác chưa được gán nhãn được thu thập từ Internet. Thực tế cho thấy rất nhiều các bài toán Machine Learning thuộc vào nhóm này vì việc thu thập dữ liệu có nhãn tốn rất nhiều thời gian và có chi phí cao. Rất nhiều loại dữ liệu thậm chí cần phải có chuyên gia mới gán nhãn được (ảnh y học chẳng hạn). Ngược lại, dữ liệu chưa có nhãn có thể được thu thập với chi phí thấp từ Internet.

2.4. Học tăng cường (Reinforcement Learning)

Trong kiểu học này, các thuật toán cố gắng dự đoán đầu ra cho một vấn đề dựa trên

một tập các tham số điều chỉnh. Sau đó, đầu ra được tính toán trở thành một tham số đầu vào và đầu ra mới được tính cho đến khi thấy đầu ra tối ưu. Các thuật toán học máy được sử dụng cho kiểu học này như: Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) và Deep Learning. Học tăng cường chủ yếu được sử dụng trong phát triển các ứng dụng giải trí AI, thu nhận kỹ năng, điều hướng robot và quyết định thời gian thực.

3. Một số ứng dụng của Học máy

Hầu hết mọi ngành công nghiệp đang làm việc với hàm lượng lớn dữ liệu đều nhận ra tầm quan trọng của công nghệ Học máy. Những thấu hiểu một cách sâu sắc (insights) từ nguồn dữ liệu này sẽ giúp các tổ chức vận hành hiệu quả hơn hoặc tạo được lợi thế cạnh tranh so với các đối thủ.

- **Các dịch vụ tài chính:** Ngân hàng và những doanh nghiệp hoạt động trong lĩnh vực tài chính sử dụng các mô hình Học máy với hai mục đích chính: xác định insights trong dữ liệu và ngăn chặn lừa đảo. Insights sẽ biết được các cơ hội đầu tư hoặc thông báo đến nhà đầu tư thời điểm giao dịch hợp lý. Học máy cũng có thể tìm được những khách hàng đang có hồ sơ rủi ro cao hoặc sử dụng giám sát mạng để chỉ rõ những tín hiệu lừa đảo.

- **Chính phủ:** Các tổ chức Chính phủ hoạt động về an ninh cộng đồng hoặc tiện ích xã hội sở hữu rất nhiều nguồn dữ liệu có thể khai thác insights. Ví dụ, khi phân tích dữ liệu cảm biến, chính phủ sẽ tăng mức độ hiệu quả của dịch vụ và tiết kiệm chi phí. Học máy còn hỗ trợ phát hiện gian lận và giảm thiểu khả năng trộm cắp danh tính.

- **Chăm sóc sức khỏe:** Học máy là một xu hướng phát triển nhanh chóng trong ngành chăm sóc sức khỏe, nhờ vào sự ra đời của các thiết bị và máy cảm ứng đeo được sử dụng dữ liệu để đánh giá tình hình sức khỏe của bệnh nhân trong thời gian thực (real-time). Học máy còn giúp các chuyên gia y tế xác định những xu hướng hoặc tín hiệu để cải thiện khả năng điều trị, chuẩn đoán bệnh.

- **Marketing và bán hàng:** Dựa trên hành vi mua hàng trước đây, các trang web sử dụng Học máy phân tích lịch sử mua hàng, từ đó giới thiệu những vật dụng mà bạn có thể sẽ quan tâm và yêu thích. Khả năng tiếp nhận dữ liệu, phân tích và sử dụng những dữ liệu đó để cá nhân hóa trải nghiệm mua sắm (hoặc thực hiện chiến dịch Marketing) chính là tương lai của ngành bán lẻ.

- **Dầu khí:** Học máy giúp tìm kiếm những nguồn nguyên liệu mới, phân tích các mỏ dầu dưới đất, dự đoán tình trạng thất bại của bộ cảm biến lọc dầu hay sắp xếp các kênh phân phối để đạt hiệu quả và tiết kiệm chi phí. Có thể nói, số lượng các trường hợp sử dụng Học máy trong ngành công nghiệp này cực kì lớn và vẫn ngày càng mở rộng

- **Vận tải:** Phân tích dữ liệu để xác định mẫu (patterns) và các xu hướng là trọng tâm trong ngành vận tải vì đây là ngành phụ thuộc vào khả năng tận dụng hiệu quả trên mỗi tuyến đường và dự đoán các vấn đề tiềm tàng để gia tăng lợi nhuận. Các chức năng phân tích dữ liệu và mô hình của Học máy đóng vai trò quan trọng với các doanh nghiệp vận chuyển, vận tải công cộng và các tổ chức vận chuyển khác.

Ngoài ra, Học máy đã giúp cho công nghệ xe tự hành phát triển mạnh và đã được nhiều tập đoàn công nghệ lớn, xe hơi như Uber, Tesla, Apple, BMW, GM Cruise... đang thử nghiệm và bước đầu triển khai tại thị trường Bắc Mỹ.

4. Kết luận

Có thể nói, Học máy đã và đang phủ khắp mọi lĩnh vực trong cuộc sống hiện tại và tương lai. Nhờ có Học máy, các công việc đã được tự động hóa và trở nên dễ dàng hơn rất nhiều. Các máy móc tưởng chừng như vô tri vô giác nhưng cũng có thể có khả năng học như con người và chúng ta đang dần biến chúng có những trí thông minh, giác quan như con người. Việc học của máy tính là cần phải có dữ liệu, dữ liệu càng nhiều thì máy tính càng được học nhiều và hiệu quả càng cao. Đó là lý do mặc dù Học máy được giới thiệu từ cuối những năm 1950 nhưng cho tới thời gian gần đây Học máy mới thực

sự phát triển mạnh nhờ có những kho dữ liệu lớn để học. Ngược lại, với dữ liệu lớn thì các công cụ xử lý dữ liệu truyền thống gặp rất nhiều khó khăn và thường không đạt hiệu quả cao, do đó việc sử dụng các kỹ thuật học máy thực hiện trích xuất thông tin tự động từ dữ liệu để cho ra những thông tin hữu ích là một nhiệm vụ cấp bách hơn bao giờ hết.

Tài liệu tham khảo

1. Đặng Thị Hằng (2019), *Trí tuệ nhân tạo, Máy học, Dữ liệu lớn*, truy cập ngày 02/04/2019 từ <https://www.phamduytung.com/blog/2019-04-02-deep-learning-view/>.
2. Phạm Ngọc Thọ (2019), *Vai trò của học máy trong phân tích dữ liệu trong các ứng dụng IoT*, truy cập ngày 29/11/2019 từ <http://csnd.vn/Home/Giao-duc-Dao-tao/6640/Vai-tro-cua-hoc-may-trong-phan-tich-du-lieu-trong-cac-ung-dung-IoT>.
3. Vũ Hữu Tiệp (2018), *Machine Learning cơ bản*, ebook.