FACULTY OF TECHNOLOGY AND DATA SCIENCE

HỘI NGHỊ KHOA HỌC

NGHIÊN CỨU HƯỚNG TỚI PHÁT TRIỂN NGÀNH ĐÀO TẠO KHOA HỌC MÁY TÍNH VÀ DỮ LIỆU TRONG KINH TẾ VÀ KINH DOANH CỦA TRƯỜNG ĐH NGOẠI THƯƠNG

HÀ NỘI, 6 - 2024

LỜI MỞ ĐẦU

Trong không khí toàn thể cán bộ, giảng viên của Nhà trường đang gấp rút hoàn thành các công việc kết thúc năm học 2023 – 2024 và chuẩn bị đón chào kỳ hè 2024, Khoa Công nghệ và Khoa học dữ liệu tổ chức hội nghị khoa học cấp Khoa thường niên: **"NGHIÊN CỨU HƯỚNG TỚI PHÁT TRIỂN NGÀNH ĐÀO TẠO KHOA HỌC MÁY TÍNH VÀ DỮ LIỆU TRONG KINH TẾ VÀ KINH DOANH CỦA TRƯỜNG ĐẠI HỌC NGOẠI THƯƠNG"**. Các bài viết của hội nghị lần này tập trung vào nội dung chính: *Giới thiệu nội dung các phương pháp giảng dạy hiện đại, các hướng nghiên cứu đã và đang được triển khai thành công đối với các học phần Khoa học Máy tính và dữ liệu, Toán, Tin nhằm nâng cao năng lực giảng dạy và nghiên cứu khoa học của giảng viên bộ môn Toán, Tin và bộ môn Khoa học máy tính và dữ liệu để phục vụ phát triển các ngành Khoa học máy tính, chương trình đào tạo Khoa học máy tính và dữ liệu trong kinh tế & kinh doanh , các bài viết giới thiệu các kết quả nghiên cứu về khoa học Công nghệ, Toán và Tin ứng dụng để giải quyết một số bài toán thực tế.*

Trong quá trình chuẩn bị Hội thảo, Ban tổ chức Hội thảo không chỉ nhận được bài viết của các giảng viên trong khoa mà còn nhận được các bài viết của các giảng viên Cơ sở 2; các trường đại học bạn. Ban tổ chức Hội nghị xin nhiệt liệt biểu dương và đánh giá cao sự tham gia nhiệt tình của các nhà khoa học trong quá trình chuẩn bị bài viết, biên tập bài viết cho Hội nghị cũng như quá trình chuẩn bị các mặt để tổ chức Hội nghị này. Ban tổ chức Hội nghị rất mong nhận được các ý kiến đóng góp quý báu từ quý vị đại biểu và các độc giả.

Nhân dịp này, Khoa Công nghệ và Khoa học dữ liệu chân thành gửi lời cảm ơn tới Đảng Ủy, Ban Giám Hiệu Nhà trường, Phòng Quản lý Khoa học, Phòng Quản lý Đào tạo, Phòng Kế hoạch Tài chính cùng các đơn vị trong toàn trường về sự ủng hộ, giúp đỡ, động viên các cán bộ, giảng viên Khoa Công nghệ và Khoa học dữ liệu trong công tác nghiên cứu khoa học.

Xin chân thành cảm ơn các đại biểu đã tới dự và tham gia thảo luận đóng góp cho sự thành công của hội nghị.

T/M Ban chủ nhiệm Khoa
P.Trưởng Khoa
ThS Tô Thị Hải Yến

DANH MỤC BÀI VIẾT

PHẦN 1. KHOA HỌC MÁY TÍNH VÀ PHÂN TÍCH DỮ LIỆU

PHẦN 2: CÁC MÔ HÌNH PHÂN TÍCH ĐỊNH LƯỢNG VÀ ỨNG DỤNG

TẦM QUAN TRỌNG CỦA DỮ LIỆU TRONG HỌC MÁY VÀ CÁC CHIẾN LƯỢC ĐỂ XÂY DỰNG BỘ DỮ LIỆU TỐT

: TS. Lê Bích Phượng, TS. Nguyễn Thị Hằng
Khoa Khoa học Cơ bản, Trường Đại học Mỏ - Địa chất, 18 Phố Viên, Hà Nội
Email lebichphuong@humg.edu.vn; Tel: 0988782112

Tóm tắt: *Dữ liệu đóng vai trò quan trọng trong học máy, ảnh hưởng trực tiếp đến hiệu suất của mô hình và khả năng tổng quát hóa của nó. Bài báo nhấn mạnh về tầm quan trọng của dữ liệu trong học máy và cung cấp các chiến lược để xây dựng dữ liệu tốt nhằm tối ưu hóa hiệu suất của các mô hình học máy, bao gồm:Thu thập dữ liệu đa dạng từ nhiều nguồn khác nhau để đảm bảo tính đại diện cho mọi khía cạnh của vấn đề;làm sạch và tiền xử lý dữ liệu để loại bỏ nhiễu và chuẩn hóa định dạng, tăng khả năng học của mô hình, xác minh tính chính xác của dữ liệu và tính đúng đắn của nhãn (nếu có) để đảm bảo tính đáng tin cậy của mô hình, sử dụng kỹ thuật tạo dữ liệu tổng hợp để bổ sung dữ liệu hiện có, đặc biệt là trong trường hợp dữ liệu quá ít hoặc không đủ đa dạng. Những nỗ lực này không chỉ cải thiện hiệu suất của mô hình mà còn tạo ra những ứng dụng học máy đáng tin cậy và có thể áp dụng trong thực tế.*

Từ khóa: dữ liệu, học máy, hiệu suất mô hình

1. Giới thiệu chung

Học máy là một lĩnh vực của trí tuệ nhân tạo mà máy tính được lập trình để tự động học và cải thiện hiệu suất từ dữ liệu mà không cần phải được lập trình một cách tường tận. Có hai dạng chính của học máy:

Học máy giám sát (Supervised Learning): Mô hình được đào tạo trên một tập dữ liệu gồm các cặp (input, output), và mục tiêu là dự đoán đầu ra cho các input mới. Ví dụ, trong việc phân loại email là rác hay không, bạn cung cấp cho mô hình các email đã được đánh dấu là rác hoặc không rác, và sau đó mô hình có thể dự đoán xem một email mới có phải là rác hay không. Học máy không giám sát (Unsupervised Learning): Mô hình được đào tạo trên dữ liệu mà không có nhãn, và mục tiêu là tìm ra cấu trúc hoặc mẫu ẩn trong dữ liệu. Ví dụ, trong việc gom nhóm dữ liệu, mô hình có thể tự động phân chia dữ liệu thành các nhóm dựa trên sự tương đồng của chúng. [1, 2, 3].

Ngoài ra, còn có các phương pháp khác như học tăng cường (reinforcement learning), học bán giám sát (semi-supervised learning), và học sâu (deep learning), nơi mà các mạng nơ-ron sâu được sử dụng để học các biểu diễn phức tạp từ dữ liệu. Học máy có nhiều ứng dụng trong nhiều lĩnh vực như xử lý ngôn ngữ tự nhiên, thị giác máy tính, dự đoán, và nhiều lĩnh vực khác [4, 5].

Trong học máy, có một số yếu tố quan trọng mà ảnh hưởng đến hiệu suất của mô hình và quá trình học như: Dữ liệu, thuật toán, kiến trúc mô hình, tham số, thuật toán tối ưu, hàm thất thoát.

1. Dữ liệu: Dữ liệu là yếu tố quan trọng nhất trong học máy. Chất lượng và đa dạng của dữ liệu đều ảnh hưởng đến khả năng học của mô hình. Dữ liệu phải

được làm sạch và chuẩn bị cẩn thận để đảm bảo rằng mô hình có thể học được từ nó một cách hiệu quả.

2. Thuật toán: Thuật toán được sử dụng để đào tạo mô hình cũng rất quan trọng. Có nhiều loại thuật toán khác nhau được sử dụng cho các vấn đề khác nhau trong học máy, và việc chọn đúng thuật toán phù hợp với bài toán cụ thể là rất quan trọng.

3. Kiến trúc mô hình: Kiến trúc của mô hình đóng vai trò quan trọng trong việc xác định khả năng học và khả năng tổng quát hóa của mô hình. Đối với các bài toán khác nhau, có thể cần sử dụng các kiến trúc mô hình khác nhau để đạt được hiệu suất tốt nhất.

4. Tham số và siêu tham số: Các tham số và siêu tham số của mô hình cũng ảnh hưởng đến hiệu suất của nó. Việc điều chỉnh các tham số này một cách cẩn thận có thể cải thiện hiệu suất của mô hình.

5. Tối ưu hóa: Thuật toán tối ưu hóa được sử dụng để điều chỉnh các tham số của mô hình dựa trên dữ liệu đào tạo. Việc chọn đúng thuật toán tối ưu hóa và điều chỉnh các siêu tham số của nó là rất quan trọng để đạt được hiệu suất tốt nhất của mô hình.

6. Đánh giá hiệu suất: Việc đánh giá hiệu suất của mô hình một cách chính xác và đáng tin cậy là rất quan trọng để hiểu được khả năng tổng quát hóa của nó và để có thể so sánh với các mô hình khác.

Tất cả những yếu tố này cùng đóng góp vào việc xây dựng một mô hình học máy hiệu quả và có khả năng tổng quát hóa tốt trên dữ liệu mới.

Dữ liệu đóng vai trò cực kỳ quan trọng trong quá trình học máy, ảnh hưởng trực tiếp đến hiệu suất của mô hình và khả năng tổng quát hóa của nó. Bài báo này nhấn mạnh về tầm quan trọng của dữ liệu trong học máy và cung cấp các chiến lược để xây dựng dữ liệu tốt nhằm tối ưu hóa hiệu suất của các mô hình học máy.

2. Nội dung

2.1. Sự quan trọng của Dữ liệu:

Dữ liệu là nguyên liệu cơ bản trong quá trình học máy, nó quyết định đến hiệu suất và khả năng tổng quát hóa của mô hình cũng như khả năng áp dụng của nó trong các tình huống thực tế. Chất lượng của dữ liệu ảnh hưởng trực tiếp đến khả năng học của mô hình. Dữ liệu không chính xác, không đủ đa dạng hoặc không đại diện cho thực tế có thể dẫn đến mô hình học máy kém chính xác và không thể tổng quát hóa trên dữ liệu mới. Chúng ta có thể thấy rõ hơn qua các ví dụ sau:

- Nhận diện ảnh khuôn mặt: Một hệ thống nhận diện khuôn mặt được đào tạo chỉ trên dữ liệu từ một nhóm dân tộc nhất định sẽ thiếu khả năng nhận diện khuôn mặt của các cá nhân khác dân tộc. Điều này làm giảm tính tổng quát hóa của mô hình và gây ra các vấn đề về công bằng và đa dạng dân tộc trong ứng dụng thực tế.

- Dự đoán thời tiết: Nếu dữ liệu về thời tiết chỉ được thu thập từ một số khu vực nhất định mà không đại diện cho các điều kiện khí hậu đa dạng, mô hình dự đoán thời tiết có thể bị thiên vị và không chính xác khi áp dụng cho các vùng lân cận hoặc các vùng địa lý khác.

- Tự động gợi ý sản phẩm: Một hệ thống gợi ý sản phẩm dựa trên dữ liệu mua sắm có thể bị hạn chế nếu chỉ tập trung vào một số nhóm sản phẩm cụ thể, bỏ qua các sở thích và nhu cầu tiêu dùng của các nhóm khách hàng khác.
- Dự đoán khả năng trả nợ của khách hàng: Nếu dữ liệu về lịch sử tín dụng không đầy đủ hoặc không chính xác, mô hình dự đoán khả năng trả nợ có thể không đáng tin cậy, dẫn đến quyết định về tín dụng không chính xác.
- Phát hiện gian lận tín dụng: Dữ liệu không chính xác hoặc không đầy đủ về các giao dịch gian lận có thể khiến mô hình phát hiện gian lận tín dụng không hiệu quả, ảnh hưởng đến tính toàn vẹn của hệ thống tài chính.
- Tư vấn y tế: Nếu dữ liệu y tế không đại diện cho một loạt các điều kiện sức khỏe và đối tượng bệnh nhân, mô hình tư vấn y tế có thể đưa ra lời khuyên không phù hợp cho các nhóm đối tượng bệnh nhân khác nhau.
- Tạo phân loại hình ảnh xe tự lái: Dữ liệu chỉ từ một số điều kiện giao thông hoặc một số khu vực địa lý có thể dẫn đến việc mô hình không nhận diện được các tình huống giao thông đặc biệt hoặc không chuẩn bị cho các điều kiện giao thông khác nhau.
- Dự đoán giá nhà: Dữ liệu không chính xác hoặc không đầy đủ về giá nhà có thể dẫn đến mô hình dự đoán giá nhà không chính xác, ảnh hưởng đến quyết định mua bán và đầu tư trong lĩnh vực bất động sản.
- Tự động phát hiện spam email: Nếu dữ liệu chỉ bao gồm các loại spam cũ và không đa dạng, mô hình phát hiện spam có thể bỏ qua các loại spam mới và tinh vi, gây ra nguy cơ cho hộp thư đến của người dùng.
- Dự đoán kết quả trận đấu thể thao: Nếu dữ liệu về kết quả trận đấu không đủ hoặc không đáng tin cậy, mô hình dự đoán kết quả có thể không tin cậy, không giúp ích cho các nhà cái hoặc người chơi đặt cược.

2.2. Các chiến lược để xây dựng bộ dữ liệu tốt

- Thu thập dữ liệu đa dạng: Thu thập dữ liệu từ nhiều nguồn và nguồn gốc khác nhau để đảm bảo tính đại diện cho mọi khía cạnh của vấn đề. Điều này giúp mô hình họ Trong việc đoán bệnh ung thư da qua ảnh, việc thu thập dữ liệu đa dạng từ nhiều nguồn có thể giúp cải thiện hiệu suất của mô hình học máy. Dưới đây là một số ví dụ về việc thu thập dữ liệu đa dạng trong lĩnh vực này:

+) Thu thập ảnh từ nhiều nguồn khác nhau: Bao gồm ảnh chụp từ các bệnh viện và phòng khám da liễu khác nhau trên toàn thế giới. Các bệnh viện ở các vùng địa lý khác nhau có thể gặp phải các loại ung thư da đặc trưng riêng, và việc thu thập dữ liệu từ các nguồn khác nhau sẽ giúp mô hình học máy hiểu được sự biến động này.

+) Sử dụng các công nghệ mới để thu thập dữ liệu: Ngoài việc sử dụng ảnh chụp từ các thiết bị y tế truyền thống, cũng có thể sử dụng các công nghệ mới như hình ảnh từ drone hoặc camera di động để thu thập dữ liệu từ các môi trường ngoài trời hoặc khó tiếp cận.

+) Thu thập ảnh từ nhiều loại thiết bị và góc chụp: Đảm bảo rằng dữ liệu bao gồm ảnh chụp từ nhiều loại thiết bị (ví dụ: máy ảnh chuyên nghiệp, điện thoại di động) và từ nhiều góc độ khác nhau. Điều này giúp mô hình học máy có thể học được sự biến đổi của bệnh trên nhiều điều kiện ánh sáng và góc chụp.

+) Thu thập ảnh từ các trường hợp khác nhau của bệnh: Bệnh ung thư da có thể biến đổi rất nhiều ở mỗi người, từ kích thước, hình dạng đến màu sắc. Việc thu thập ảnh từ các trường hợp khác nhau, bao gồm cả các trường hợp bệnh lý nặng và nhẹ, giúp mô hình học máy có thể đào tạo được tốt hơn.c máy hiểu được các biến thể và tình huống khác nhau.

- Làm sạch và tiền xử lý dữ liệu: Trước khi đưa vào mô hình học máy, dữ liệu cần phải được làm sạch và tiền xử lý để loại bỏ nhiễu và chuẩn hóa định dạng. Điều này giúp loại bỏ thông tin không cần thiết và tăng khả năng học của mô hình. Dưới đây là một số ví dụ về các lệnh làm sạch và tiền xử lý dữ liệu trong Excel và Python:

+) Trong Excel:

1. Loại bỏ dòng trống và dữ liệu trùng lặp:
   - Sử dụng tính năng lọc để loại bỏ dòng trống.
   - Sử dụng chức năng Remove Duplicates để loại bỏ các dòng có dữ liệu trùng lặp.
2. Chuẩn hóa dữ liệu:
   - Sử dụng chức năng Text to Columns để tách dữ liệu trong một ô thành nhiều cột dựa trên một ký tự phân tách nhất định.
   - Sử dụng các công thức như UPPER, LOWER hoặc PROPER để chuẩn hóa chữ viết hoa, chữ viết thường hoặc chữ hoa chữ cái đầu.
3. Xử lý dữ liệu thiếu:
   - Sử dụng chức năng Fill Down để điền các ô trống bằng giá trị của ô phía trên hoặc phía dưới.
   - Sử dụng tính năng Replace để thay thế các ô trống bằng giá trị được chỉ định.

Trong Python (sử dụng thư viện pandas):

1. Loại bỏ dòng trống và dữ liệu trùng lặp:

```
# Loại bỏ dòng trống
df.dropna(inplace=True)
# Loại bỏ dữ liệu trùng lặp
df.drop_duplicates(inplace=True)
```

2. Chuẩn hóa dữ liệu:

```
# Chuẩn hóa chữ viết hoa
df['Column'] = df['Column'].str.upper()
# Chuẩn hóa chữ viết thường
df['Column'] = df['Column'].str.lower()
```

3. Xử lý dữ liệu thiếu:

```
# Điền các giá trị trống bằng giá trị trung bình của cột
df.fillna(df.mean(), inplace=True)
# Loại bỏ các hàng chứa dữ liệu trống
df.dropna(inplace=True)
```

Những lệnh này giúp làm sạch và tiền xử lý dữ liệu một cách hiệu quả trước khi đưa vào mô hình học máy.

- Xác minh và chuẩn xác dữ liệu: Kiểm tra tính chính xác của dữ liệu và xác minh tính đúng đắn của nhãn (nếu có). Điều này giúp tránh các lỗi trong quá trình học và đảm bảo tính đáng tin cậy của mô hình.
- Tạo dữ liệu tổng hợp (Synthetic Data): Sử dụng kỹ thuật tạo dữ liệu tổng hợp để bổ sung dữ liệu hiện có, đặc biệt là trong trường hợp dữ liệu hiện có quá ít hoặc không đủ đa dạng.

3.Kết luận:

Dữ liệu đóng vai trò quan trọng trong quá trình học máy và ảnh hưởng trực tiếp đến hiệu suất của mô hình. Để đạt được hiệu suất tốt nhất, cần áp dụng các chiến lược xây dựng dữ liệu tốt như thu thập dữ liệu đa dạng, làm sạch và tiền xử lý dữ liệu, xác minh và chuẩn xác dữ liệu, cũng như sử dụng dữ liệu tổng hợp khi cần thiết. Những nỗ lực này không chỉ cải thiện hiệu suất của mô hình mà còn tạo ra những ứng dụng học máy đáng tin cậy và có thể áp dụng trong thực tế.

TÀI LIỆU THAM KHẢO

[1].Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

[2].Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners.

[3].Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A. R., Levy, O., ... & Zettlemoyer, L. (2020). "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension."

[4].Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). "Exploring the limits of transfer learning with a unified text-to-text transformer."

[5].Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding."