ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CHƯƠNG 6
Công nghệ khai phá dữ liệu trong doanh nghiệp

Biên soạn: ThS. Nguyễn Thị Anh Thư

Nội dung

1. Giới thiệu
2. Phân loại thuật toán
3. Các bài toán khai phá dữ liệu
4. Học có giám sát
5. Học không giám sát
6. Học bán giám sát
7. Học tăng cường
8. Học chuyển giao
9. Tổng kết

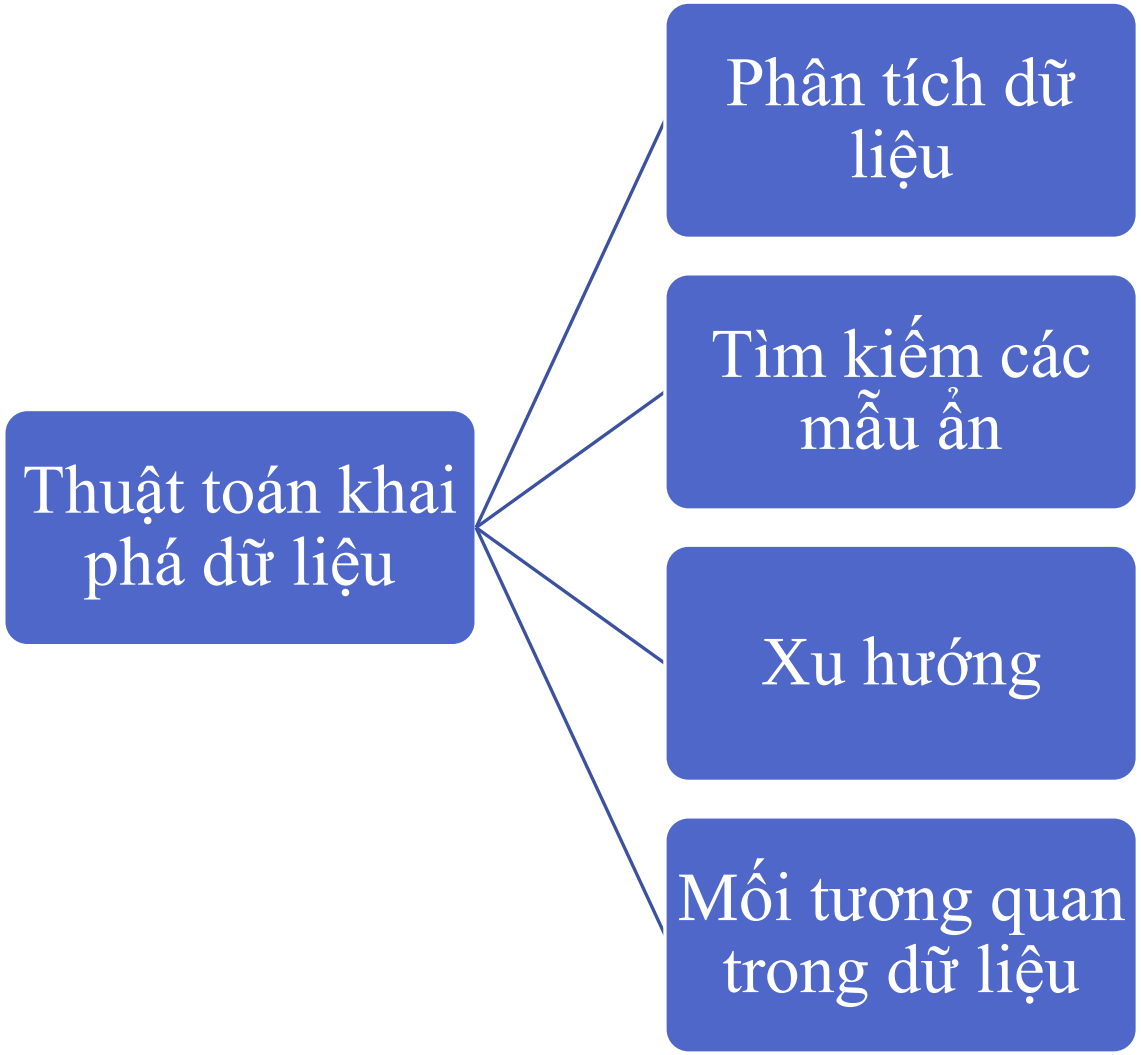
1. Giới thiệu

Data mining (Khai phá dữ liệu) là một lĩnh vực liên ngành thuộc khoa học máy tính, sử dụng các kỹ thuật thống kê, học máy và quản lý cơ sở dữ liệu để trích xuất kiến thức và thông tin hữu ích từ dữ liệu lớn.

Mục tiêu là chuyển đổi dữ liệu thô thành thông tin có giá trị có thể được sử dụng để:

- ***Đưa ra dự đoán***: Dự đoán xu hướng tương lai và đưa ra quyết định sáng suốt hơn.
- ***Khám phá tri thức***: Xác định các mẫu và mối quan hệ ẩn trong dữ liệu mà có thể không được nhìn thấy rõ ràng.
- ***Phân loại dữ liệu***: Nhóm các dữ liệu tương tự nhau vào các nhóm.
- ***Tối ưu hóa quy trình***: Xác định các yếu tố ảnh hưởng đến hiệu quả hoạt động và đề xuất các giải pháp cải tiến.

1. Giới thiệu



Khai phá dữ liệu (Data Mining) là một lĩnh vực của khoa học máy tính tập trung vào việc **trích xuất thông tin hữu ích từ dữ liệu thô**.

2. Phân loại thuật toán

Theo mục đích sử dụng

- **Học máy (Machine Learning):** Các thuật toán được sử dụng để huấn luyện mô hình từ dữ liệu và đưa ra dự đoán.
- **Xử lý ngôn ngữ tự nhiên (Natural Language Processing):** Các thuật toán được sử dụng để xử lý và hiểu ngôn ngữ con người.
- **Thị giác máy tính(Computer Vision):** Các thuật toán được sử dụng để xử lý và hiểu hình ảnh và video.
- **Khai phá dữ liệu (Data Mining):** Các thuật toán được sử dụng để trích xuất thông tin từ dữ liệu.

2. Phân loại thuật toán

Theo phương thức học

- **Học có giám sát (Supervised Learning):** Các thuật toán được huấn luyện trên tập dữ liệu đã được gắn nhãn.
- **Học không giám sát (Unsupervised Learning):** Các thuật toán được huấn luyện trên tập dữ liệu chưa được gắn nhãn.
- **Học bán giám sát (Semi-Supervised Learning):** Các thuật toán được huấn luyện trên tập dữ liệu kết hợp cả dữ liệu đã được gắn nhãn và chưa được gắn nhãn.
- **Học tăng cường (Reinforcement Learning):** Phương pháp học máy cho phép các tác nhân học cách hành động trong môi trường để tối đa hóa phần thưởng.
- **Học chuyển giao (Transfer Learning):** Kỹ thuật học máy trong đó kiến thức được học từ một bài toán được áp dụng cho một bài toán khác.

2. Phân loại thuật toán

Theo loại mô hình

- **Mạng nơ-ron nhân tạo (Artificial Neural Networks):** Các mô hình được lấy cảm hứng từ cấu trúc của não bộ con người.
- **Cây quyết định (Decision Trees):** Các mô hình sử dụng các quy tắc để đưa ra dự đoán.
- **Máy vector hỗ trợ (Support Vector Machines):** Các mô hình tìm kiếm đường phân chia tối ưu giữa các lớp dữ liệu.

2. Phân loại thuật toán

Theo độ phức tạp

- **Thuật toán đơn giản:** Các thuật toán dễ hiểu và dễ triển khai.
- **Thuật toán phức tạp:** Các thuật toán có hiệu suất cao nhưng khó hiểu và khó triển khai.

2. Phân loại thuật toán

Việc phân loại các thuật toán chỉ mang tính tương đối và có thể chồng chéo nhau.

Một số thuật toán có thể được sử dụng cho nhiều mục đích khác nhau.

Việc lựa chọn thuật toán phù hợp phụ thuộc vào nhiều yếu tố, bao gồm loại bài toán, kích thước tập dữ liệu, chất lượng dữ liệu và độ chính xác mong muốn.

3. Các bài toán khai phá dữ liệu

Thuật toán khai phá dữ liệu	Phân loại (Classification)
	Phân cụm (Clustering)
	Luật kết hợp (Association Rule)
	Hồi quy (Regression)
	Chuỗi thời gian (Time series)

Phân loại (Classification)

Học có giám sát (Supervised Learning)

Phân loại (Classification)

Được sử dụng để dự đoán nhãn (label) cho các dữ liệu mới dựa trên các dữ liệu đã được dán nhãn sẵn (training data).

Có hai loại chính:

- **Phân loại nhị phân**: Dự đoán dữ liệu thuộc một trong hai lớp (ví dụ: email spam hay không spam, khách hàng tiềm năng hay không tiềm năng).
- **Phân loại đa lớp**: Dự đoán dữ liệu thuộc một trong nhiều lớp (ví dụ: loại hoa, loại bệnh ung thư).

Phân cụm (Clustering)

Học không giám sát (Unsupervised Learning)

Phân cụm (Clustering)

Khám phá cấu trúc ẩn trong dữ liệu bằng cách nhóm các dữ liệu tương đồng vào cùng một cụm.

Mục đích:

- Tìm kiếm các mẫu (pattern) trong dữ liệu.
- Hiểu rõ hơn về cấu trúc dữ liệu.
- Phân chia dữ liệu thành các nhóm có ý nghĩa.
- Tăng hiệu quả của các thuật toán khác như phân loại, dự đoán.

Luật kết hợp (Association Rule)

Học không giám sát (Unsupervised Learning)

Luật kết hợp (Association Rule)

Tìm kiếm các mối quan hệ tương quan giữa các mục (item) trong tập dữ liệu.

Mục đích:

- Khám phá các quy tắc ẩn trong dữ liệu.
- Xác định các tập mục thường xuyên xuất hiện cùng nhau.
- Tìm kiếm các cơ hội bán hàng chéo (cross-selling) và tiếp thị (marketing).
- Phân tích hành vi khách hàng.

Hồi quy (Regression)

Học có giám sát (Supervised Learning)

Hồi quy (Regression)

Thuật toán Hồi quy sử dụng dữ liệu đã được dán nhãn để học mối quan hệ giữa biến phụ thuộc (dependent variable) và các biến độc lập (independent variables), từ đó dự đoán giá trị của biến phụ thuộc cho các dữ liệu mới.

Mục đích:

- Dự đoán giá trị của một biến trong tương lai.
- Tìm kiếm mối quan hệ giữa các biến.
- Xây dựng mô hình để giải thích các hiện tượng.

Chuỗi thời gian (Time series)

Chuỗi thời gian là một tập dữ liệu được thu thập theo thời gian, bao gồm các giá trị được đo lường tại các thời điểm khác nhau.

Thuật toán chuỗi thời gian phân tích dữ liệu chuỗi thời gian và trích xuất thông tin hữu ích từ dữ liệu.

- Có thể thuộc nhiều phương thức học khác nhau, tùy thuộc vào loại bài toán và cách thức áp dụng.

Mục đích:

- Dự đoán giá trị trong tương lai.
- Phát hiện các xu hướng và chu kỳ trong dữ liệu.
- Xác định các điểm bất thường (anomaly) trong dữ liệu.
- Hiểu rõ hơn về hành vi của hệ thống.

4. Học có giám sát

1. Dữ liệu:

- Sử dụng tập dữ liệu được gắn nhãn (labeled dataset) với cặp đầu vào (input) và đầu ra (output) mong muốn.
- Dữ liệu đầu vào có thể là hình ảnh, văn bản, âm thanh, hoặc bất kỳ dạng dữ liệu nào khác.
- Dữ liệu đầu ra có thể là giá trị liên tục (hồi quy) hoặc giá trị rời rạc (phân loại).

2. Quá trình học:

- Thuật toán học từ các ví dụ được cung cấp trong tập dữ liệu.
- Tìm ra mối quan hệ giữa đầu vào và đầu ra.
- Xây dựng mô hình dự đoán đầu ra cho dữ liệu mới.

3. Loại mô hình:

- **Phân loại (Classification):** Dự đoán nhãn cho dữ liệu đầu vào. Ví dụ: phân loại email là spam hay không spam.
- **Hồi quy (Regression):** Dự đoán giá trị liên tục cho dữ liệu đầu vào. Ví dụ: dự đoán giá nhà dựa trên diện tích và vị trí.

4. Học có giám sát

Ưu điểm

- Hiệu quả cao với dữ liệu được gắn nhãn tốt.
- Dễ dàng triển khai và sử dụng.
- Có nhiều thuật toán cho các bài toán khác nhau.
- Có thể giải thích được kết quả dự đoán.

Nhược điểm

- Phụ thuộc vào chất lượng dữ liệu.
- Cần nhiều dữ liệu để huấn luyện mô hình hiệu quả.
- Khó khăn trong việc thu thập dữ liệu được gắn nhãn.
- Mô hình có thể bị quá khớp (overfitting) hoặc thiếu khớp (underfitting) với dữ liệu.

5. Học không giám sát

1. Dữ liệu:

- Sử dụng tập dữ liệu không được gắn nhãn (unlabeled dataset).
- Dữ liệu đầu vào có thể là hình ảnh, văn bản, âm thanh, hoặc bất kỳ dạng dữ liệu nào khác.
- Dữ liệu đầu ra được dự đoán bởi thuật toán.

2. Quá trình học:

- Thuật toán tự tìm ra cấu trúc và mối quan hệ trong dữ liệu.
- Không có hướng dẫn cung cấp câu trả lời đúng.
- Mục tiêu là khám phá các mẫu ẩn trong dữ liệu.

3. Loại mô hình:

- **Phân cụm (Clustering):** Nhóm các điểm dữ liệu có đặc điểm tương tự nhau vào cùng một nhóm.
- **Giảm chiều (Dimensionality Reduction):** Giảm số lượng đặc trưng của dữ liệu mà vẫn giữ được thông tin quan trọng.
- **Phát hiện anomalya (Anomaly Detection):** Phát hiện các điểm dữ liệu khác biệt so với phần còn lại của dữ liệu.

5. Học không giám sát

Ưu điểm

- Không cần dữ liệu được gắn nhãn.
- Có thể áp dụng cho các trường hợp dữ liệu không có sẵn nhãn.
- Khám phá các mẫu ẩn trong dữ liệu.
- Có thể được sử dụng để chuẩn bị dữ liệu cho các thuật toán học có giám sát.

Nhược điểm

- Khó khăn trong việc đánh giá hiệu quả mô hình.
- Kết quả dự đoán có thể khó giải thích.
- Phụ thuộc vào lựa chọn thuật toán và tham số.

6. Học bán giám sát

1. Dữ liệu:

- Sử dụng tập dữ liệu kết hợp cả dữ liệu được gắn nhãn và dữ liệu chưa được gắn nhãn.
- Dữ liệu được gắn nhãn cung cấp thông tin cho thuật toán học.
- Dữ liệu chưa được gắn nhãn giúp cải thiện hiệu quả của mô hình.

2. Quá trình học:

- Thuật toán học từ cả dữ liệu được gắn nhãn và dữ liệu chưa được gắn nhãn.
- Tận dụng thông tin trong cả hai loại dữ liệu để xây dựng mô hình tốt hơn.
- Có nhiều thuật toán học bán giám sát khác nhau.

6. Học bán giám sát

Ưu điểm:

- Cải thiện hiệu quả mô hình với lượng dữ liệu được gắn nhãn hạn chế.
- Tiết kiệm chi phí thu thập dữ liệu được gắn nhãn.
- Có thể áp dụng cho các trường hợp dữ liệu được gắn nhãn đắt đỏ hoặc khó thu thập.

Nhược điểm:

- Khó khăn trong việc lựa chọn thuật toán phù hợp.
- Hiệu quả mô hình phụ thuộc vào chất lượng dữ liệu chưa được gắn nhãn.
- Có thể xảy ra hiện tượng nhiễu dữ liệu.

6. Học bán giám sát

Một số thuật toán học bán giám sát:

- **Self-training**: Mô hình học từ các dự đoán của chính nó.
- **Co-training**: Sử dụng hai mô hình khác nhau để học từ nhau.
- **Label propagation**: Truyền nhãn từ các ảnh được gắn nhãn sang các ảnh chưa được gắn nhãn.

6. Học bán giám sát

Self-training

Bước 1: Thu thập dữ liệu
- Thu thập một tập dữ liệu gồm dữ liệu được gắn nhãn và dữ liệu chưa được gắn nhãn.
- Gắn nhãn cho một số dữ liệu trong tập dữ liệu (ví dụ: 100 dữ liệu).
- Giữ lại các dữ liệu còn lại không gắn nhãn.

Bước 2: Huấn luyện mô hình ban đầu
- Sử dụng dữ liệu được gắn nhãn để huấn luyện một mô hình ban đầu.

Bước 3: Dự đoán nhãn cho dữ liệu chưa được gắn nhãn
- Sử dụng mô hình ban đầu để dự đoán nhãn cho các dữ liệu chưa được gắn nhãn. Các dự đoán này được sử dụng để cập nhật mô hình, giúp cải thiện hiệu quả của mô hình.

Bước 4: Cập nhật mô hình
- Sử dụng các dự đoán từ bước 3 để cập nhật mô hình.

Bước 5: Lặp lại các bước 3 và 4
- Lặp lại các bước 3 và 4 cho đến khi đạt được hiệu quả mong muốn.

6. Học bán giám sát

Self-training

Bước 1: Thu thập dữ liệu
- Thu thập một tập dữ liệu gồm 1000 ảnh mèo và 1000 ảnh chó.
- Gắn nhãn cho 100 ảnh mèo và 100 ảnh chó.
- Giữ lại 800 ảnh mèo và 800 ảnh chó chưa được gắn nhãn.

Bước 2: Huấn luyện mô hình ban đầu
- Sử dụng 100 ảnh mèo và 100 ảnh chó được gắn nhãn để huấn luyện một mô hình ban đầu.

Bước 3: Dự đoán nhãn cho dữ liệu chưa được gắn nhãn
- Sử dụng mô hình ban đầu để dự đoán nhãn cho 800 ảnh mèo và 800 ảnh chó chưa được gắn nhãn.

Bước 4: Cập nhật mô hình
- Sử dụng các dự đoán từ bước 3 để cập nhật mô hình.

Bước 5: Lặp lại các bước 3 và 4
- Lặp lại các bước 3 và 4 cho đến khi đạt được hiệu quả mong muốn.

7. Học tăng cường

1. Khái niệm:

- Học tăng cường là một lĩnh vực của học máy cho phép agent (hệ thống) học cách hành động trong môi trường để *tối đa hóa phần thưởng nhận được*. Agent tự khám phá môi trường, thử nghiệm các hành động khác nhau và học hỏi từ những tương tác với môi trường đó.

2. Các thành phần chính:

- **Agent**: Hệ thống học tập và thực hiện hành động trong môi trường.
- **Môi trường**: Cung cấp thông tin cho agent và phản hồi lại các hành động của agent.
- **Hành động**: Các lựa chọn mà agent có thể thực hiện trong môi trường.
- **Trạng thái**: Mô tả tình trạng hiện tại của môi trường.
- **Phần thưởng**: Giá trị phản hồi mức độ tốt của hành động mà agent thực hiện.

7. Học tăng cường

3. Quá trình học:

- **Khám phá**: Agent khám phá môi trường và thử nghiệm các hành động khác nhau.
- **Tương tác**: Agent nhận thông tin từ môi trường và cập nhật trạng thái hiện tại.
- **Học tập**: Agent học cách liên kết hành động với phần thưởng.
- **Lập kế hoạch**: Agent sử dụng kiến thức đã học để lựa chọn hành động tối ưu trong tương lai.

4. Loại hình học tăng cường:

- **Học tăng cường dựa trên giá trị**: Agent học giá trị của từng trạng thái hoặc hành động.
- **Học tăng cường dựa trên chính sách**: Agent học chính sách trực tiếp, là ánh xạ từ trạng thái sang hành động.

7. Học tăng cường

Ưu điểm:

- Có thể giải quyết các bài toán phức tạp trong môi trường không xác định.
- Cho phép agent tự học hỏi và thích ứng với môi trường.
- Có thể áp dụng cho nhiều lĩnh vực khác nhau.

Nhược điểm:

- Khó khăn trong việc thiết kế môi trường và phần thưởng.
- Quá trình học tập có thể diễn ra chậm chạp.
- Có thể gặp vấn đề với tính ổn định và hiệu quả.

7. Học tăng cường

Q-learning là một thuật toán học tăng cường dựa trên giá trị. Thuật toán này sử dụng hàm Q để biểu thị giá trị dự kiến của việc thực hiện một hành động trong một trạng thái nhất định.

Hàm Q là một bảng 2 chiều, với:

- **Trục X**: Các *trạng thái có thể xảy* ra trong môi trường.
- **Trục Y**: Các *hành động có thể thực hiện* trong mỗi trạng thái.

Giá trị **Q(s, a)** biểu thị giá trị dự kiến của việc thực hiện ***hành động a trong trạng thái s***.

7. Học tăng cường

Q-learning

Có thể được áp dụng để cải thiện hiệu quả của chatbot. Chatbot sử dụng Q-learning để học cách tương tác với người dùng và đạt được mục tiêu mong muốn.

Giả sử chúng ta muốn xây dựng một chatbot để hỗ trợ khách hàng mua sắm trực tuyến. Chatbot có thể thực hiện các hành động sau:

- **Hỏi thông tin về sản phẩm**: Chatbot hỏi khách hàng về sản phẩm họ quan tâm.
- **Giới thiệu sản phẩm**: Chatbot giới thiệu các sản phẩm phù hợp với nhu cầu của khách hàng.
- **Trả lời câu hỏi**: Chatbot trả lời các câu hỏi của khách hàng về sản phẩm.
- **Hoàn tất đơn hàng**: Chatbot giúp khách hàng hoàn tất đơn hàng.

7. Học tăng cường

Q-learning

Quá trình học:

- Khởi tạo: Khởi tạo hàm Q với giá trị ngẫu nhiên cho tất cả trạng thái và hành động.
- Lặp lại:
  - ***Chọn hành động***: Chatbot chọn hành động có giá trị Q cao nhất trong trạng thái hiện tại.
  - ***Thực hiện hành động***: Chatbot thực hiện hành động được chọn.
  - ***Nhận phản hồi***: Chatbot nhận phần thưởng từ người dùng nếu hành động giúp người dùng đạt được mục tiêu và bị phạt nếu hành động không hiệu quả.
  - ***Cập nhật hàm Q***: Chatbot cập nhật giá trị Q cho hành động đã thực hiện dựa trên phần thưởng nhận được và giá trị Q cao nhất cho các hành động có thể thực hiện trong trạng thái tiếp theo.

8. Học chuyển giao

1. Khái niệm:

- Học chuyển giao là một kỹ thuật học máy cho phép sử dụng kiến thức học được từ một bài toán (nhiệm vụ nguồn) để giải quyết một bài toán khác (nhiệm vụ mục tiêu).

2. Cách thức hoạt động:

- **Huấn luyện mô hình trên nhiệm vụ nguồn**: Mô hình được huấn luyện trên một tập dữ liệu lớn và có sẵn cho nhiệm vụ nguồn.
- **Chuyển giao kiến thức**: Kiến thức học được từ mô hình nhiệm vụ nguồn được chuyển giao sang mô hình nhiệm vụ mục tiêu.
- **Huấn luyện mô hình trên nhiệm vụ mục tiêu**: Mô hình nhiệm vụ mục tiêu được huấn luyện trên một tập dữ liệu nhỏ hơn cho nhiệm vụ mục tiêu.

8. Học chuyển giao

Ưu điểm:

- Cải thiện hiệu quả mô hình với lượng dữ liệu ít cho nhiệm vụ mục tiêu.
- Tiết kiệm thời gian và chi phí thu thập dữ liệu.
- Có thể áp dụng cho các trường hợp dữ liệu cho nhiệm vụ mục tiêu đắt đỏ hoặc khó thu thập.

Nhược điểm:

- Khó khăn trong việc lựa chọn mô hình nguồn phù hợp.
- Hiệu quả mô hình phụ thuộc vào sự tương đồng giữa hai nhiệm vụ.
- Có thể xảy ra hiện tượng nhiễu dữ liệu từ mô hình nguồn.

8. Học chuyển giao

Học chuyển giao BERT là sử dụng mô hình BERT được huấn luyện trên một tập dữ liệu lớn (nhiệm vụ nguồn) để giải quyết một bài toán khác (nhiệm vụ mục tiêu).

Cách thức hoạt động:

- **Bước 1**: Huấn luyện mô hình BERT trên tập dữ liệu nguồn.
- **Bước 2**: Chuyển giao kiến thức từ mô hình BERT sang mô hình nhiệm vụ mục tiêu.
- **Bước 3**: Huấn luyện mô hình nhiệm vụ mục tiêu với lượng dữ liệu nhỏ hơn.

8. Học chuyển giao

Ví dụ:

- **Nhiệm vụ nguồn**: Phân loại câu tweet là tích cực hay tiêu cực.
- **Tập dữ liệu nguồn**: 1 triệu câu tweet với nhãn tích cực hoặc tiêu cực.
- **Nhiệm vụ mục tiêu**: Phân loại câu trả lời trên forum là tích cực hay tiêu cực.
- **Tập dữ liệu mục tiêu**: 10.000 câu trả lời trên forum với nhãn tích cực hoặc tiêu cực.

Cách thức thực hiện:

- Bước 1: Huấn luyện mô hình BERT trên tập dữ liệu tweet.
- Bước 2: Chuyển giao kiến thức từ mô hình BERT sang mô hình phân loại câu trả lời forum.
- Bước 3: Huấn luyện mô hình phân loại câu trả lời forum với 10.000 câu trả lời trên forum.

Kết quả: Mô hình phân loại câu trả lời forum được huấn luyện bằng phương pháp học chuyển giao BERT đạt hiệu quả cao hơn so với mô hình được huấn luyện từ đầu.

9. Tổng kết

Thuật toán	Mô tả	Ưu điểm	Nhược điểm	Ví dụ
Học có giám sát	Học từ tập dữ liệu có nhãn	Hiệu quả cao, dễ hiểu	Yêu cầu dữ liệu có nhãn	Phân loại ảnh, dự đoán giá nhà
Học không giám sát	Học từ tập dữ liệu không nhãn	Khám phá cấu trúc ẩn trong dữ liệu	Khó đánh giá hiệu quả	Phân cụm khách hàng, phát hiện gian lận
Học bán giám sát	Học từ tập dữ liệu có và không nhãn	Cải thiện hiệu quả với lượng dữ liệu ít	Phụ thuộc vào chất lượng dữ liệu có nhãn	Phân loại văn bản, dự đoán bệnh
Học tăng cường	Học từ môi trường bằng cách thử nghiệm	Khả năng thích ứng cao	Khó khăn trong thiết kế môi trường	Chơi game, điều khiển robot
Học chuyển giao	Tận dụng kiến thức từ một bài toán để giải quyết bài toán khác	Tiết kiệm thời gian và chi phí	Hiệu quả phụ thuộc vào sự tương đồng giữa hai bài toán	Xác định cảm xúc, dịch máy

Question & Answer