KIỂM TRA KHAI THÁC DỮ LIỆU

khai thác dữ liệu và ứng dụng (Trường Đại học Công nghiệp Thành phố Hồ Chí Minh)

KIỂM TRA KHAI THÁC DỮ LIỆU

CHƯƠNG 1:

Khai thác dữ liệu là gì?

Khai thác dữ liệu (data mining) là quá trình khám phá thông tin có giá trị từ các tập dữ liệu lớn để tìm ra các mẫu, quy luật, và mối quan hệ không được biết trước đến từ dữ liệu đó. Đây là một phần quan trọng của khoa học dữ liệu và phân tích dữ liệu, nó thường được áp dụng trong nhiều lĩnh vực như kinh doanh, y tế, khoa học xã hội, và công nghệ thông tin.

Lợi ích của bài toán khai thác dữ liệu:

Khám phá thông tin mới: Khai thác dữ liệu giúp phát hiện các mẫu, quy luật và thông tin mới từ các tập dữ liệu lớn mà trước đó không được biết đến. Điều này có thể giúp tổ chức hiểu sâu hơn về hoạt động của họ và thị trường mà họ hoạt động.

Dự đoán và tiên đoán: Bằng cách phân tích dữ liệu lịch sử, khai thác dữ liệu có thể dùng để dự đoán xu hướng tương lai và sự kiện tiềm ẩn. Điều này hữu ích cho việc ra quyết định chiến lược và kế hoạch tương lai.

Tối ưu hóa quy trình: Bằng cách phân tích dữ liệu về quy trình hoạt động, tổ chức có thể tìm ra các cách để tối ưu hóa hiệu suất và giảm thiểu lãng phí. Điều này có thể áp dụng cho nhiều lĩnh vực, từ sản xuất đến dịch vụ khách hàng.

Phát hiện gian lận và rủi ro: Khai thác dữ liệu có thể giúp phát hiện các hoạt động gian lận và rủi ro bằng cách xác định các mẫu không bình thường hoặc hành vi đáng ngờ trong dữ liệu.

Phân tích khách hàng và thị trường: Bằng cách phân tích dữ liệu khách hàng và thị trường, tổ chức có thể hiểu rõ hơn về nhu cầu, sở thích và hành vi của khách hàng, từ đó tạo ra các chiến lược tiếp thị và dịch vụ tốt hơn.

Hỗ trợ ra quyết định: Khai thác dữ liệu cung cấp thông tin chính xác và đáng tin cậy để hỗ trợ quyết định kinh doanh và chiến lược tổ chức.

Trình bày các nhiệm vụ - ứng dụng cụ thể của bài toán khai thác dữ liệu:

Phân tích hành vi khách hàng:

Dự đoán hành vi mua hàng của khách hàng dựa trên lịch sử mua hàng và thông tin cá nhân.

Phân loại khách hàng thành các nhóm dựa trên các đặc điểm chung để tạo ra chiến lược tiếp thị hiệu quả.

Phát hiện sớm các biểu hiện của sự chuyển đổi hoặc mất khách hàng để thực hiện các biện pháp giữ chân khách hàng.

Chẩn đoán y tế:

Dự đoán nguy cơ mắc các bệnh lý dựa trên dữ liệu lịch sử y tế và yếu tố rủi ro.

Phân tích các mẫu và xu hướng trong dữ liệu y tế để phát hiện ra các bệnh lý hiếm gặp hoặc tổn thương không rõ nguyên nhân.

Tối ưu hóa việc chẩn đoán và điều trị bằng cách áp dụng thông tin từ dữ liệu lâm sàng và y học.

Dự đoán nhu cầu và dự trữ hàng tồn kho:

Dự đoán nhu cầu hàng tồn kho trong tương lai dựa trên mẫu mua hàng và thị trường.

Tối ưu hóa lượng tồn kho và quản lý chuỗi cung ứng bằng cách phân tích dữ liệu về nhu cầu và mua sắm.

Phân tích dữ liệu tài chính:

Dự đoán xu hướng thị trường tài chính và đầu tư dựa trên dữ liệu lịch sử và chỉ số kinh tế.

Phát hiện gian lận tài chính và giao dịch không bình thường bằng cách xác định các mẫu không đối xứng và biến thể.

Tối ưu hóa hoạt động sản xuất:

Dự đoán nhu cầu sản phẩm và vật liệu nguyên liệu để tối ưu hóa quá trình sản xuất và lập kế hoạch sản xuất.

Phát hiện ra các lỗi sản xuất và hỏng hóc thiết bị bằng cách theo dõi các dữ liệu cảm biến và quá trình sản xuất.

Phân tích dữ liệu xã hội và truyền thông:

Phân tích ý kiến và tư duy của người dùng trên mạng xã hội để đo lường sự hài lòng khách hàng và phản ứng của công chúng đối với sản phẩm hoặc dịch vụ.

Dự đoán xu hướng truyền thông và viral của các nội dung trên mạng xã hội để tối ưu hóa chiến lược tiếp thị và quảng cáo.

CHƯƠNG 2: Tiền xử lý dữ liệu

Xử lý dữ liệu khử nhiều với chuyển dữ liệu khử nhiều:

+ Phương pháp rời rạc hóa dữ liệu : trình bày các bước theo chiều rộng/chiều sâu

+ Áp dụng các bước đó trên tập dữ liệu nhỏ để xử lý

Biến đổi dữ liệu, chuẩn hóa dữ liệu

Chiều Rộng:

1. **Chọn các biến cần rời rạc hóa:** Xác định các biến trong tập dữ liệu mà bạn muốn biến đổi thành dạng rời rạc.
2. **Xác định các phép biến đổi cần thiết:** Xác định phép biến đổi cần áp dụng cho từng biến, như chia thành các khoảng giá trị hoặc gán nhãn.
3. **Thiết lập các ngưỡng hoặc quy tắc:** Xác định các ngưỡng hoặc quy tắc để ánh xạ giá trị của các biến thành các nhóm rời rạc.
4. **Áp dụng biến đổi:** Áp dụng phép biến đổi đã xác định cho từng biến trong tập dữ liệu.
5. **Kiểm tra và đánh giá kết quả:** Kiểm tra kết quả của quá trình rời rạc hóa để đảm bảo rằng nó đáp ứng yêu cầu và mục tiêu ban đầu.

Chiều Sâu:

1. **Chọn một biến cần rời rạc hóa:** Lựa chọn một biến cụ thể trong tập dữ liệu để thực hiện rời rạc hóa.
2. **Phân tích sâu hơn về biến đó:** Nắm vững các thuộc tính của biến, như phân phối, giá trị cực đại và cực tiểu, và các xu hướng hoặc mẫu.
3. **Chọn phép biến đổi cụ thể:** Dựa trên phân tích sâu hơn, chọn phép biến đổi phù hợp như chia thành các khoảng giá trị hoặc gán nhãn.
4. **Áp dụng phép biến đổi và kiểm tra:** Áp dụng phép biến đổi đã chọn và kiểm tra kết quả để đảm bảo rằng chúng hợp lý và phù hợp với mục tiêu phân tích.

Áp dụng các bước trên tập dữ liệu nhỏ để xử lý biến đổi dữ liệu và chuẩn hóa dữ liệu:

1. **Chọn tập dữ liệu:** Lựa chọn một tập dữ liệu nhỏ từ tập dữ liệu lớn để áp dụng các bước xử lý và rời rạc hóa.

2. **Thực hiện xử lý**: Áp dụng các bước xử lý và rời rạc hóa đã mô tả trên tập dữ liệu nhỏ.
3. **Đánh giá kết quả:** Đánh giá kết quả của quá trình xử lý và rời rạc hóa trên tập dữ liệu nhỏ để xác định hiệu suất và hiệu quả của các phương pháp được áp dụng.
4. **Tinh chỉnh và cải thiện:** Dựa trên kết quả đánh giá, điều chỉnh và cải thiện các bước xử lý và rời rạc hóa nếu cần thiết để đạt được kết quả tốt nhất trên tập dữ liệu nhỏ này.

CHƯƠNG 3: Khai thác luật kết hợp (1 trong các thuật toán)

- Trình bày thuật toán Apriori tìm tập mục phổ biến
- Trình bày thuật toán FPGrowth tìm tập mục phổ biến
- Trình bày thuật toán để xây dựng cây FP
- Trình bày thuật toán để sinh luật kết hợp từ tập mục phổ biế

Có bài toán áp dụng

+ Tìm tập mục phổ biến

+ Tìm tập mục phổ biến tối đại

+ Tìm tập mục phổ biến đóng

+ Sinh luật từ các tập mục phổ biến

+ Sinh luật từ các tập mục phổ biến tối đại

+ Sinh luật từ các tập mục phổ biến đóng

CÂU MỞ RỘNG

Cho .... xác định độ đo thú vị của luật nào đó tính được độ đo thú vị