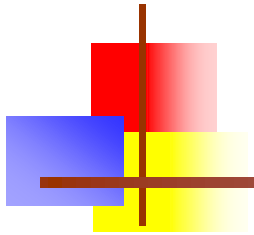


Khai phá dữ liệu (Data Mining)

Nội dung

1. Data mining là gì
2. Các vấn đề kinh doanh có thể giải quyết bằng khai phá dữ liệu
3. Các tác vụ của khai phá dữ liệu
4. Các kỹ thuật khai phá dữ liệu
5. Quy trình khai phá dữ liệu

1. Data mining là gì

- *Khai phá dữ liệu là quá trình xác định các mẫu tiềm ẩn có tính hợp lệ, mới, có ích và có thể hiểu được trong một tập hợp dữ liệu lớn một cách tự động hoặc bán tự động.*
- Một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn.
- Là những kỹ thuật khai phá tri thức tiềm ẩn có giá trị trong CSDL (Knowledge Discovery in Databases) [1989, Fayyad]
- Trích rút tri thức (Knowledge extraction).
- Phân tích mẫu dữ liệu (Data pattern analysis).

Sự cần thiết khai phá dữ liệu

- Dữ liệu lớn
  - Khai phá các mẫu tiềm ẩn có ích.
- Tăng cạnh tranh
- Công nghệ thông tin phát triển mạnh
  - Công nghệ máy tính ngày càng phát triển có thể xử lý dữ liệu ngày càng phức tạp
  - Mô hình, công cụ phát triển ứng dụng ngày càng tốt hơn -> xây dựng các ứng dụng datamining tốt hơn, chính xác hơn

Một số lĩnh vực quan trọng

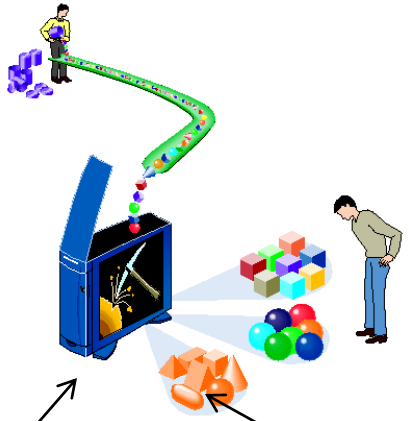
- Tài chính, ngân hàng
- Thương mại điện tử
- Bán hàng
- Tin- sinh học
- …

Các lĩnh vực nghiên cứu

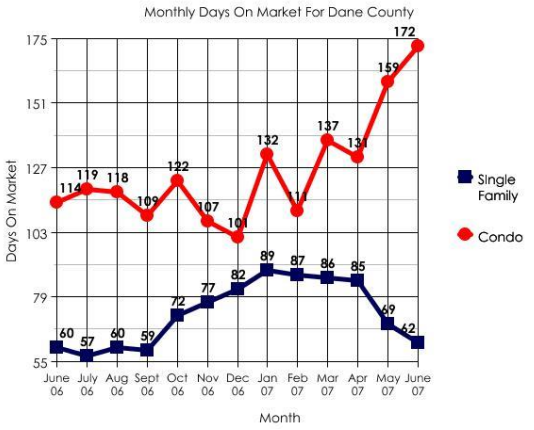
Xác suất thống kê

Trí tuệ nhân tạo

Data Mining



Tổ chức dữ liệu



Khai phá dữ liệu và tìm kiếm thông tin

- **Phân biệt rõ giữa khai phá dữ liệu với:**
  - Tìm kiếm thông tin (Information retrieval).
  - Xử lý các câu truy vấn (SQL) đối với các cơ sở dữ liệu.

Các vấn đề kinh doanh có thể giải quyết bằng khai phá dữ liệu

- Phân tích khách hàng
  - Những khách hàng nào có nhiều khả năng chuyển sang là khách hàng của đối thủ cạnh tranh??
- Bán hàng chéo (Cross –selling)
  - Những sản phẩm nào là khách hàng có thể mua sau khi đã mua một sản phẩm nào đó??
- Phát hiện gian lận
  - Khai phá dữ liệu có thể giúp xác định những yêu cầu mà nhiều khả năng là sai
- Quản lý rủi ro

Các vấn đề kinh doanh có thể giải quyết bằng khai phá dữ liệu

- Phân loại khách hàng
  - Phân loại khách hàng giúp các nhà quản lý tiếp thị hiểu được các cấu hình khác nhau của khách hàng và có những hành động tiếp thị phù hợp dựa trên các phân loại
- Mục tiêu quảng cáo
  - Quảng cáo sẽ hiển thị điều gì cho từng khách hàng truy cập cụ thể?
- Dự báo bán hang
  - Khai phá dữ liệu dự báo có thể được sử dụng để trả lời những câu hỏi dự báo liên quan đến thời gian.

Các tác vụ khai phá dữ liệu

- Có 2 chức năng chính
  - Mô tả (description):
  - Dự đoán (prediction).
- Chia thành các nhóm tác vụ sau:
  - Phân lớp (Classification)
  - Phân cụm (Clustering)
  - Hồi quy (Regression).
  - Kết hợp (Association).
  - Dự báo(Forecasting)
  - Phân tích chuỗi tuần tự (Sequence Analysis)
  - Phân tích độ lệch (Deviation Analysis).

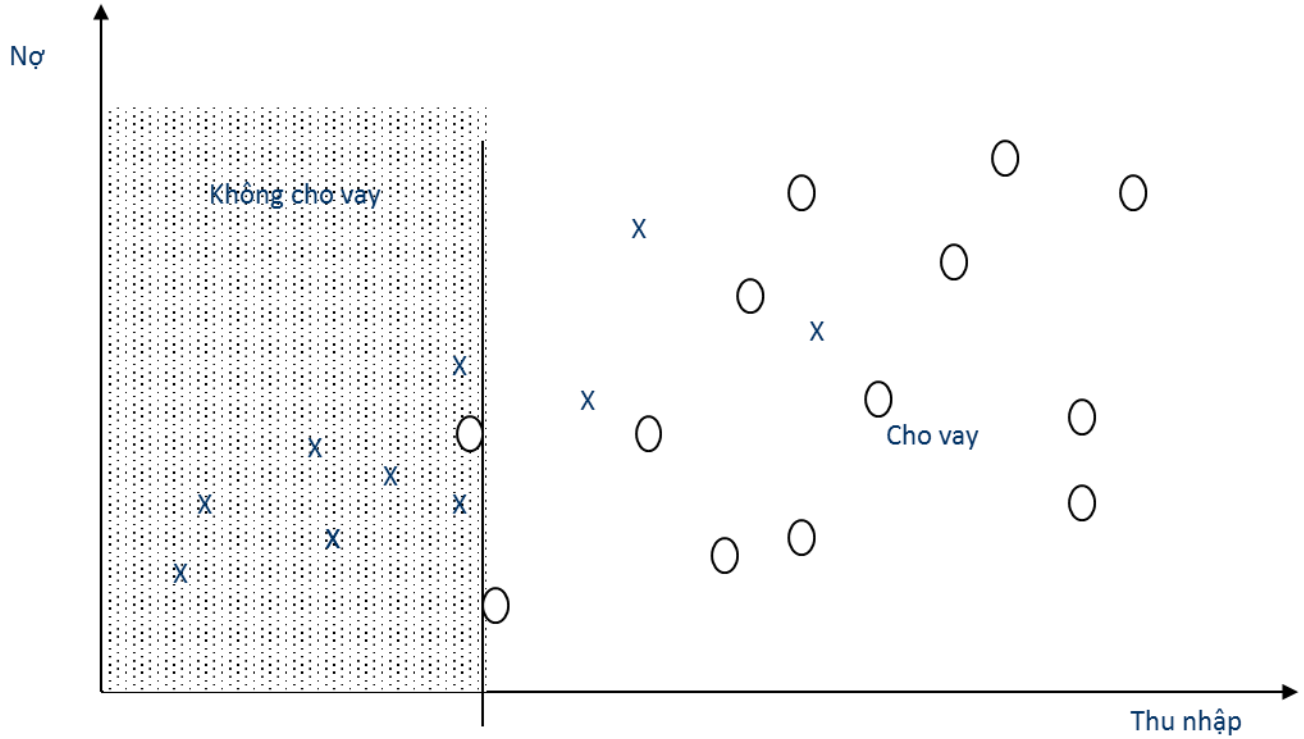
Phân lớp (Classification)

- Phân lớp (hoặc phân loại) là một trong những nhiệm vụ khai phá dữ liệu phổ biến nhất.
- Phân lớp là việc phân chia đối tượng dữ liệu vào các loại khác nhau dựa trên thuộc tính dự đoán. Mỗi đối tượng dữ liệu chứa 1 tập các thuộc tính trong đó có các thuộc tính phân lớp (thuộc tính dự đoán).
- Nhiệm vụ là tìm kiếm một mô hình mô tả các thuộc tính phân lớp (class) dữ liệu.
- Phân lớp dữ liệu được coi là quá trình học "có giám sát" (supervised), sau khi được xây dựng, mô hình phân lớp có thể được sử dụng để phân lớp các dữ liệu mới.

Phân lớp (Classification)

- Một số thuật toán:
  - Cây quyết định (decision trees)
  - Mạng nơ ron
  - Naïve Bayes
- Ví dụ: Xét một tập dữ liệu khách hàng, mỗi điểm biểu diễn 1 khách hàng đã vay của ngân hàng.Trục hoành biểu thị cho thu nhập, trục tung biểu thị cho tổng dư nợ của khách hàng. Dữ liệu khách hàng được chia thành hai lớp: dấu x biểu thị cho khách hàng bị vỡ nợ, dấu o biểu thị cho khách hàng có khả năng trả nợ.

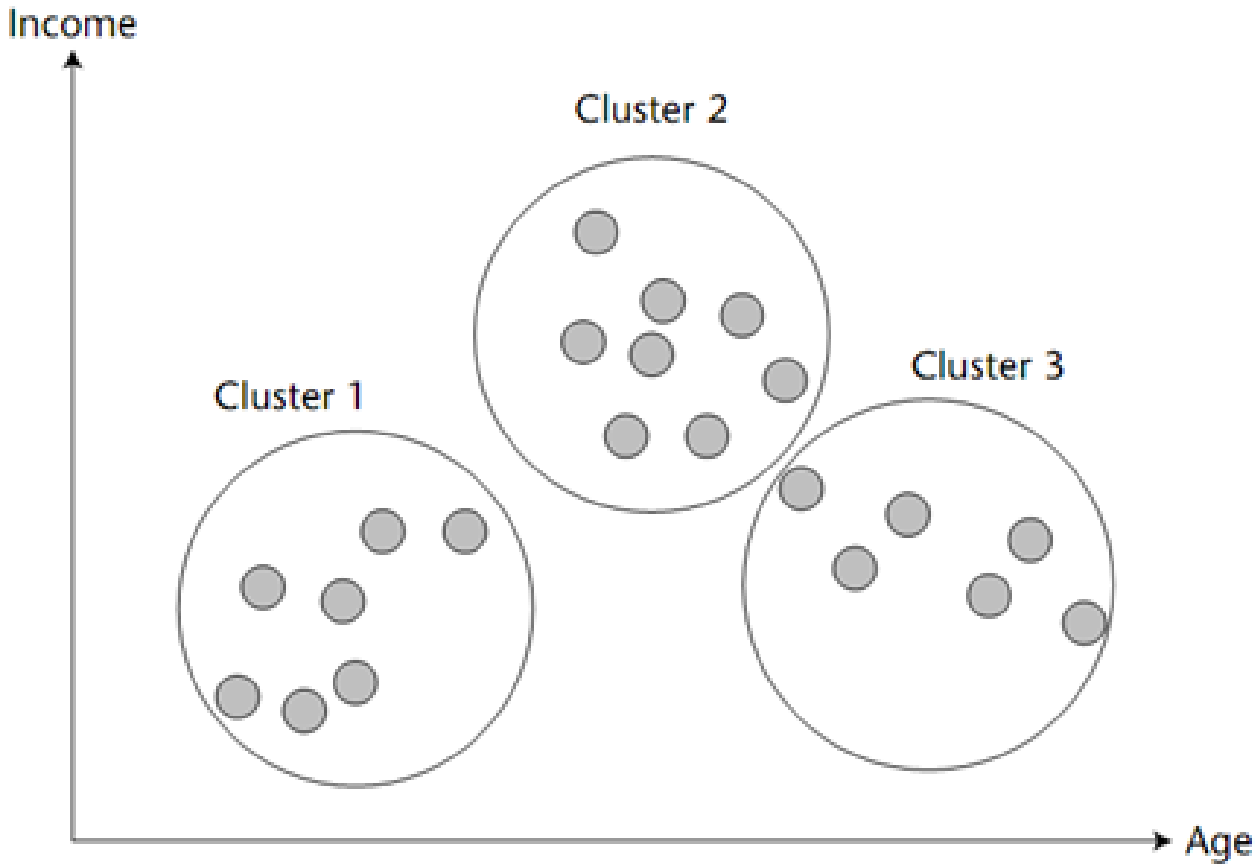
Phân lớp (Classification)



Phân cụm (Clustering)

- Phân cụm là các quy trình nhóm các đối tượng dữ liệu đã cho vào thành các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm thì tương tự(similar) nhau và các đối tượng khác cụm thì không hoặc ít tương tự hơn.
- Phân cụm được sử dụng để xác định các nhóm tự nhiên của đối tượng dữ liệu dựa trên một tập các thuộc tính.
- Phân cụm là một tác vụ khai phá dữ liệu không giám sát (unsupervised) trong học máy (Machine Learning).
  - Không có thuộc tính duy nhất được sử dụng suốt quá trình xây dựng cụm.
  - Tất cả thuộc tính đầu vào đều bình đẳng.

Phân cụm (Clustering)



Kết hợp (Association)

- Là một nhiệm vụ khai phá dữ liệu phổ biến.
- Mục tiêu: Tìm ra các mối liên hệ giữa các thành phần dữ liệu trong CSDL.
- Nhiệm vụ chính: xác định
  - Tập thuộc tính (itemset) : dựa trên ngưỡng hỗ trợ được định nghĩa bởi người sử dụng.
  - Tập luật kết hợp (*association rules*).
- Ví dụ : Trong cơ sở dữ liệu bán sản phẩm
  - ItemSet: {Product = “Pepsi”, Product = “Chips”, Product = “Juice”}.
  - {Product = “Pepsi”, Product = “Chips”} => {Product = “Juice”} với xác suất là 80%.

Hồi quy (Regression)

- Tương tự như phân lớp dữ liệu.
- Sự khác biệt chính: thuộc tính dự đoán là một số liên tục.
- Các kỹ thuật hồi quy đã được sử dụng trong lĩnh vực thống kê.
- Kỹ thuật hồi quy: Cây hồi quy và mạng nơ ron.
- Hồi quy được sử dụng để dự báo nhiều vấn đề trong kinh doanh
  - **Dự đoán lãi suất, phương pháp và khối lượng phân phối hàng hóa. Dự đoán vận tốc gió dựa trên nhiệt độ, áp suất không khí và độ ẩm, …**

Dự báo(Forecasting)

- Dự báo đóng vai trò quan trọng với khai phá dữ liệu.
- Đầu vào thường là tập dữ liệu theo chuỗi thời gian.
- Kỹ thuật dự báo thường giải quyết các vấn đề có xu hướng chung, có tính chu kỳ và lọc nhiễu.
- Dự báo giúp trả lời các câu hỏi:
  - Doanh số bán hàng sẽ nhận được trong tháng tiếp theo là bao nhiêu??
  - Giá trị cổ phiếu ngày hôm sau là bao nhiêu??

Dự báo(Forecasting)

Phân tích chuỗi tuần tự (Sequence Analysis)

- Phân tích chuỗi trình tự được sử dụng để tìm kiếm các mô hình trong một chuỗi dữ liệu rời rạc.
- Chuỗi trình tự bao gồm một loạt các giá trị rời rạc (hoặc trạng thái). Ví dụ, một khách hàng đầu tiên mua một máy tính, sau đó loa, và cuối cùng là một Webcam.
- Sự khác biệt giữa mô hình chuỗi tuần tự và chuỗi thời gian:
  - Chuỗi trình tự chứa các trạng thái rời rạc.
  - Chuỗi thời gian có chứa các số liên tục.

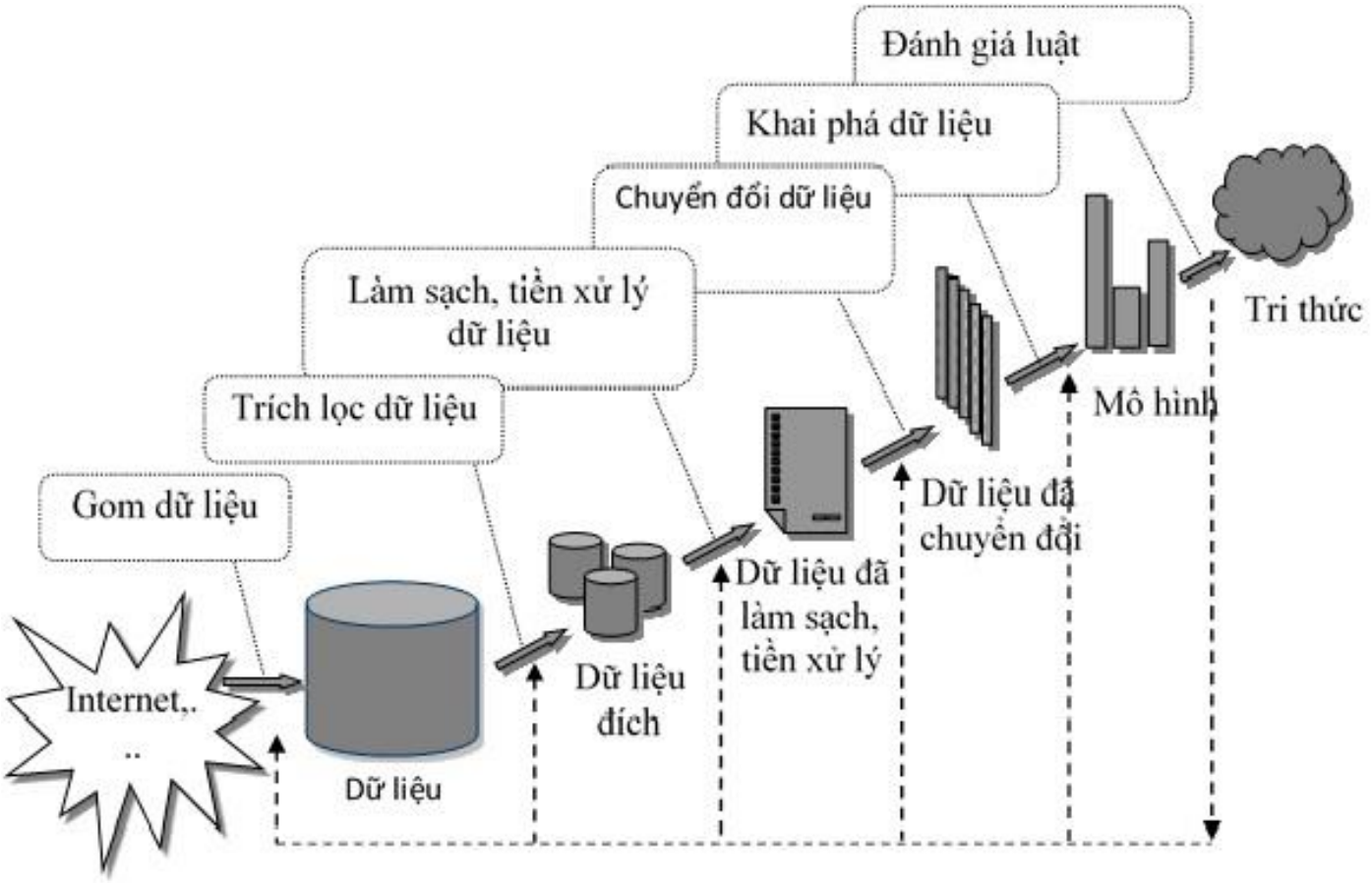
Phân tích độ lệch (Deviation Analysis)

- Phân tích độ lệch tập trung vào khai thác những thay đổi đáng kể nhất của dữ liệu so các giá trị chuẩn hoặc được đo trước đó. Nó còn được gọi là phát hiện ngoại lai.
- Không có kỹ thuật chuẩn để phân tích ngoại lai.
- Sử dụng các phiên bản sửa đổi của các thuật toán: cây quyết định, phân nhóm, hoặc các thuật toán mạng Nơron,…

Các kỹ thuật khai phá dữ liệu

- Một số thuật toán khai thác dữ liệu dựa theo thống kê:
  - Hồi quy, chuỗi thời gian, và cây quyết định.
- Thuật toán học máy (machine learning)
  - Mạng nơron (Neural networks)
  - Giải thuật di truyền

Vòng đời khai phá dữ liệu



Vòng đời khai phá dữ liệu

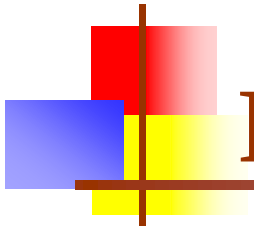
- Bước 1: Thu thập, gom dữ liệu
  - Dữ liệu được gom từ một hoặc nhiều nguồn cơ sở dữ liệu, kho dữ liệu, thậm chí dữ liệu từ những nguồn cung ứng web.
- Bước 2: Trích lọc dữ liệu
  - Dữ liệu được lựa chọn và phân chia theo một số tiêu chuẩn nào đó.
- Bước 3: Tiền xử lý dữ liệu (Làm sạch dữ liệu)
  - Loại bỏ các dữ liệu nhiễu, không đầy đủ, không nhất quán.
  - Rút gọn dữ liệu để tăng tốc độ thực hiện (gộp nhóm, lấy mẫu dữ liệu), rời rạc hóa dữ liệu…
  - Kết quả dữ liệu nhất quán, dạng chuẩn ít dư thừa và có tính đặc trưng.

Vòng đời khai phá dữ liệu

- Bước 4: Chuyển đổi dữ liệu (tranformation)
  - Biến đổi, tổ chức lại dữ liệu.
  - Chuẩn hóa dữ liệu.
  - Tạo ra dữ liệu phù hợp với mục đích khai phá dữ liệu.
  - Tùy theo thuật toán đưa về dạng phù hợp.
- Bước 5: Khai phá dữ liệu
  - Hiểu được mục tiêu khai phá và loại tác vụ khai thác dữ liệu.
  - Xây dựng các mô hình sử dụng nhiều thuật toán khác nhau và sau đó so sánh tính chính xác của các mô hình này.
  - Đưa dữ liệu vào để dự đoán kết quả mong muốn.
  - Lặp lại các bước trước nếu cần thiết.

Vòng đời khai phá dữ liệu

- Bước 6: Biểu diễn và đánh giá kết quả
  - Triển khai các thông tin dữ liệu khai phá được thành dạng các biểu đồ, báo cáo...
  - Đưa ra các mô hình dữ liệu dễ hiểu, gần gũi với người sử dụng
  - Hỗ trợ ra quyết định dựa trên việc so sánh, đánh giá các thuật toán đảm bảo độ tin cậy.



Reference

- Books online
- Microsoft SQL Server2005 For Dummies, Published by Wiley Publishing, Inc. 111 River Street