

Chapter 3:

Data Mining

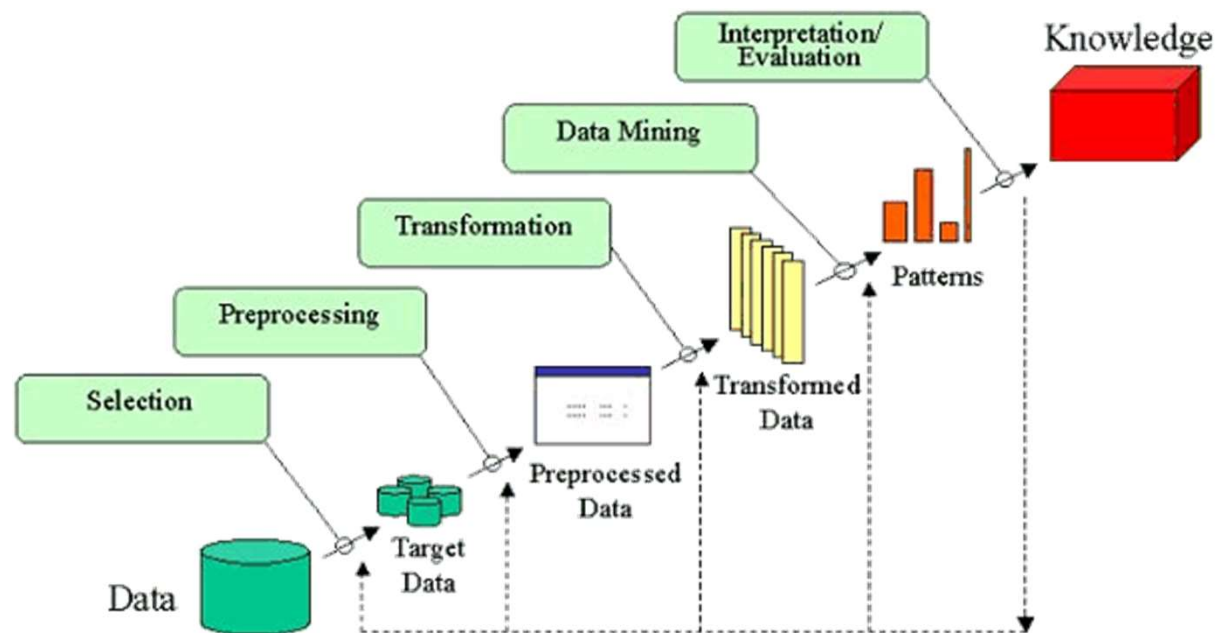
Content

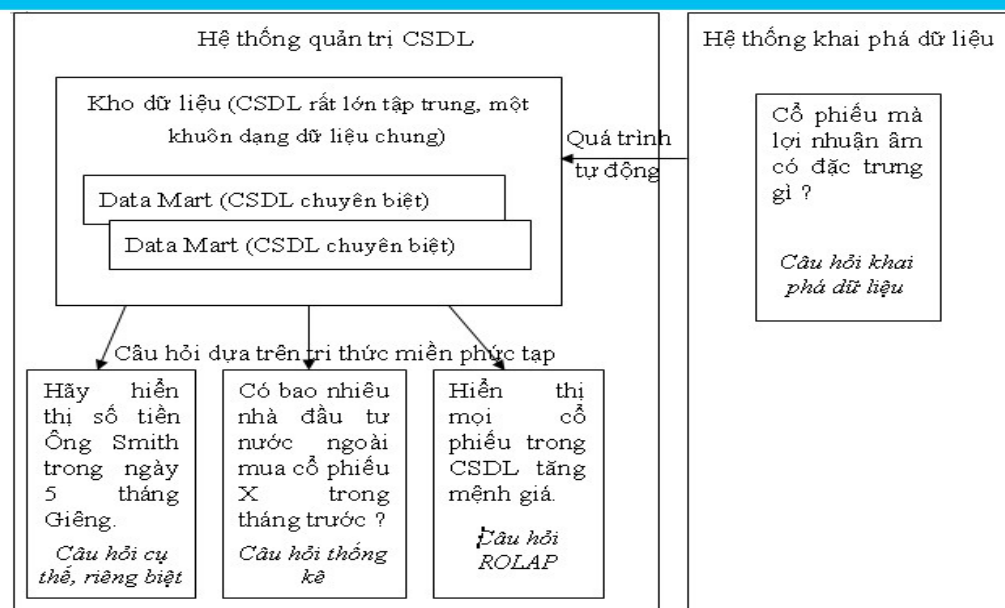
2

- » Tổng quan về Data Mining
- » Anomaly Detection
- » Association Rule Learning
- » Cluster Analysis
- » Statistical Classification
- » Regression Analysis

1. Quá trình khám phá tri thức (KDD)

3





Hình 1.7. Mối quan hệ giữa hệ thống CSDL và hệ thống khai phá dữ liệu

1. Quá trình khám phá tri thức (KDD)

5

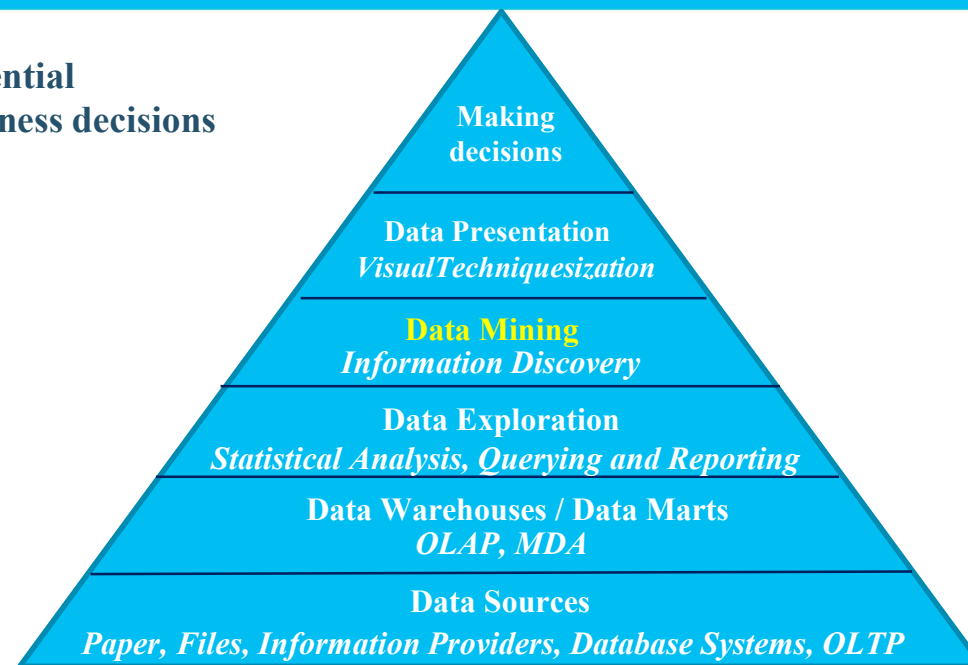
Increasing potential
to support business decisions

End User

Business
Analyst

Data
Analyst

DBA



2. Một số khái niệm cơ bản

6

Dữ liệu (**Data**): có thể xem là chuỗi các bit, là số, ký tự...mà chúng ta thu thập hàng ngày trong công việc.

Thông tin (**Information**): là tập hợp của những dữ liệu đã được xử lý, dùng mô tả, giải thích đặc tính của một đối tượng nào đó.

Tri thức (**Knowledge**): là tập hợp những thông tin có liên hệ với nhau, được lập luận chặt chẽ hoặc được thực nghiệm kiểm chứng quan nhiều thế hệ. Tri thức thể hiện tư duy của con người về một vấn đề.

2. Một số khái niệm cơ bản

7

Knowledge discovery from databases (KDD)

Khám phá tri thức từ cơ sở dữ liệu là quy trình bao gồm nhiều công đoạn như: xác định vấn đề, tập hợp và chọn lọc dữ liệu, khai thác dữ liệu, đánh giá kết quả, giải thích dữ liệu, áp dụng tri thức vào thực tế.

2. Một số khái niệm cơ bản

8

Khai phá dữ liệu (Data mining) là một bước trong quy trình khám phá tri thức, nhằm:

- Rút trích thông tin hữu ích, chưa biết, tiềm ẩn trong khối dữ liệu lớn
- Phân tích dữ liệu bán tự động
- Giải thích dữ liệu trên các tập dữ liệu lớn

2. Một số khái niệm cơ bản

9

- **Lượng lớn dữ liệu sẵn có để khai phá:**
- ❖ Bất kỳ loại dữ liệu được lưu trữ hay tạm thời, có cấu trúc hay bán cấu trúc hay phi cấu trúc
- ❖ Dữ liệu được lưu trữ
 - Các tập tin truyền thống (flat files)
 - Các cơ sở dữ liệu quan hệ (relational databases) hay quan hệ đối tượng (object relational databases)
 - Các cơ sở dữ liệu giao tác (transactional databases) hay kho dữ liệu (data warehouses)
 - Các cơ sở dữ liệu hướng ứng dụng: cơ sở dữ liệu không gian (spatial databases), cơ sở dữ liệu thời gian (temporal databases), cơ sở dữ liệu không thời gian (spatio-temporal databases), cơ sở dữ liệu chuỗi thời gian (time series databases), cơ sở dữ liệu văn bản (text databases), cơ sở dữ liệu đa phương tiện (multimedia databases), ...
 - Các kho thông tin: the World Wide Web, ...
- ❖ Dữ liệu tạm thời: các dòng dữ liệu (data streams)

2. Một số khái niệm cơ bản

10

Tri thức đạt được từ quá trình khai phá

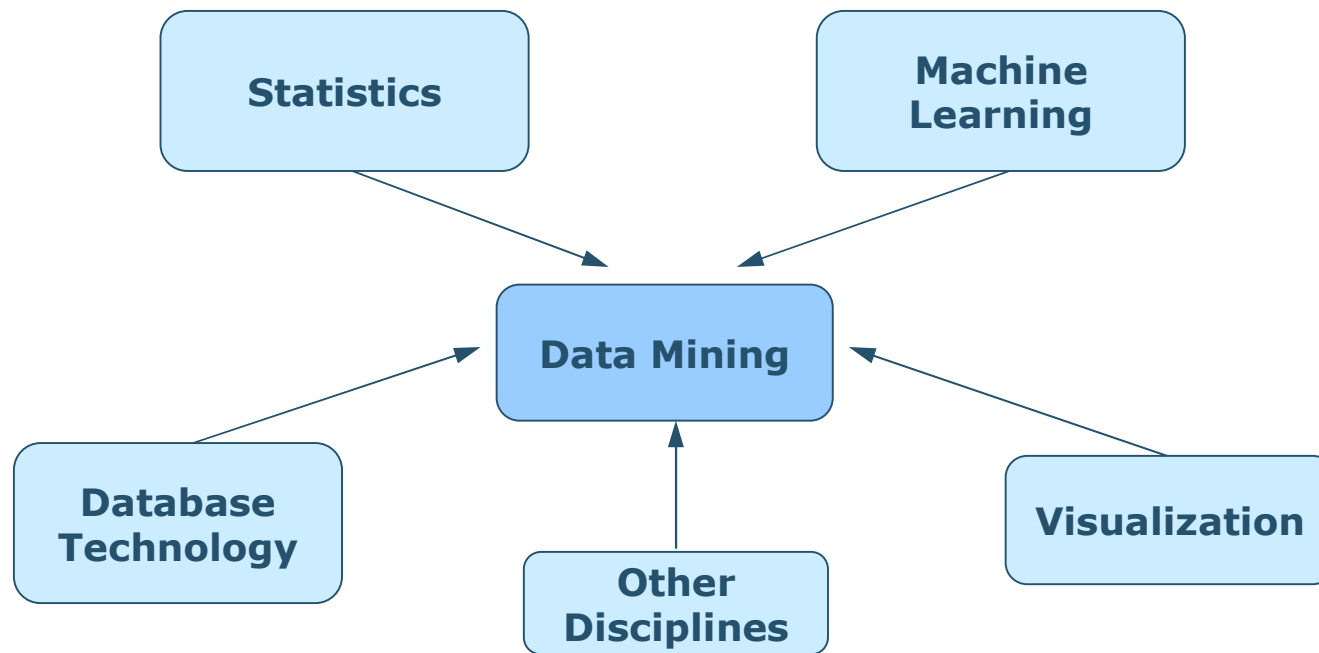
- » Mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa)
- » Mẫu thường xuyên, các mối quan hệ kết hợp/tương quan
- » Mô hình phân loại và dự đoán
- » Mô hình gom cụm
- » Các phần tử biên
- » Xu hướng hay mức độ thường xuyên của các đối tượng có hành vi thay đổi theo thời gian

2. Một số khái niệm cơ bản

11

Tri thức đạt được từ quá trình khai phá

- » Tri thức đạt được có thể có tính mô tả hay dự đoán tùy thuộc vào quá trình khai phá cụ thể.
 - » Mô tả (Descriptive): có khả năng đặc trưng hóa các thuộc tính chung của dữ liệu được khai phá (Tình huống 1)
 - » Dự đoán (Predictive): có khả năng suy luận từ dữ liệu hiện có để dự đoán (Tình huống 2, 3, và 4)
- » Tri thức đạt được có thể có cấu trúc, bán cấu trúc, hoặc phi cấu trúc.
- » Tri thức đạt được có thể được/không được người dùng quan tâm → các độ đo đánh giá tri thức đạt được.
- » Tri thức đạt được có thể được dùng trong việc hỗ trợ ra quyết định, điều khiển quy trình, quản lý thông tin, xử lý truy vấn ...



Các tác vụ khai phá dữ liệu

13

- » Khai phá mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa dữ liệu)
- » Khai phá luật kết hợp/tương quan
- » Phân loại dữ liệu
- » Dự đoán
- » Gom cụm dữ liệu
- » Phân tích xu hướng
- » Phân tích độ lệch và phân tử biên
- » Phân tích độ tương tự
- » ...

Các tác vụ khai phá dữ liệu

14

Năm thành tố cơ bản để đặc tả một tác vụ khai phá dữ liệu

- » Dữ liệu cụ thể sẽ được khai phá (task-relevant data)
- » Loại tri thức sẽ đạt được (kind of knowledge)
- » Tri thức nền (background knowledge)
- » Các độ đo (interestingness measures)
- » Các kỹ thuật biểu diễn tri thức/trực quan hóa mẫu (pattern visualization and knowledge presentation)

Các tác vụ khai phá dữ liệu

15

- » **Dữ liệu cụ thể sẽ được khai phá (task-relevant data)**
 - » Phần dữ liệu từ các dữ liệu nguồn được quan tâm
 - » Tương ứng với các thuộc tính hay chiều dữ liệu được quan tâm
 - » Bao gồm: tên kho dữ liệu/cơ sở dữ liệu, các bảng dữ liệu hay các khối dữ liệu, các điều kiện chọn dữ liệu, các thuộc tính hay chiều dữ liệu được quan tâm, các tiêu chí gom nhóm dữ liệu

Các tác vụ khai phá dữ liệu

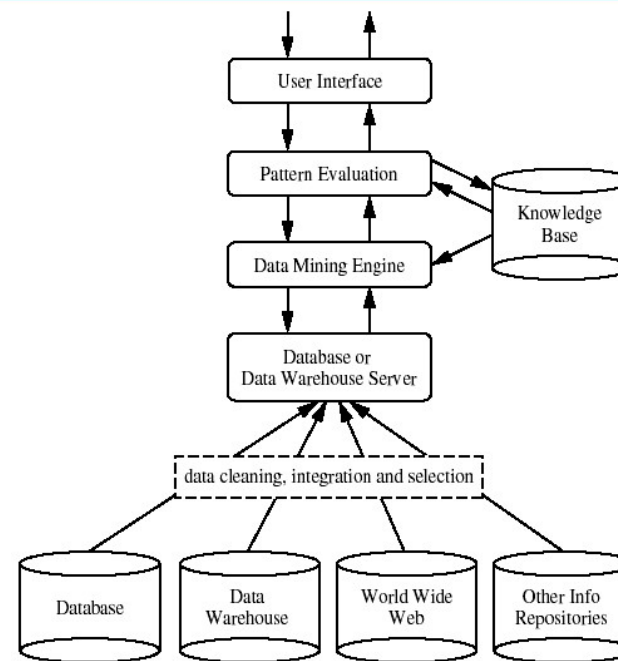
16

Loại tri thức sẽ đạt được (kind of knowledge)

- » Bao gồm: đặc trưng hóa dữ liệu, phân biệt hóa dữ liệu, mô hình phân tích kết hợp hay tương quan, mô hình phân lớp, mô hình dự đoán, mô hình gom cụm, mô hình phân tích phần tử biên, mô hình phân tích tiến hóa
- » Tương ứng với tác vụ khai phá dữ liệu cụ thể sẽ được thực thi

Kiến trúc của một hệ thống KPDL

17



Ý nghĩa và vai trò của khai phá dữ liệu

18

- » Công nghệ hiện đại trong lĩnh vực quản lý thông tin
 - » Hiện diện khắp nơi (ubiquitous) và có tính ẩn (invisible) trong nhiều khía cạnh của đời sống hằng ngày
 - » Làm việc, mua sắm, tìm kiếm thông tin, nghỉ ngơi, ...
 - » Được áp dụng trong nhiều ứng dụng thuộc nhiều lĩnh vực khác nhau
 - » Hỗ trợ các nhà khoa học, giáo dục học, kinh tế học, doanh nghiệp, khách hàng, ...

Ứng dụng của khai phá dữ liệu

19

- » Trong kinh doanh (business)
- » Trong tài chính (finance) và tiếp thị bán hàng (sales marketing)
- » Trong thương mại (commerce) và ngân hàng (bank)
- » Trong bảo hiểm (insurance)
- » Trong khoa học (science) và y sinh học (biomedicine)
- » Trong điều khiển (control) và viễn thông (telecommunication)
- » ...

Tóm tắt

20

- » Khai phá dữ liệu là quá trình khám phá ra các mẫu được quan tâm từ lượng lớn dữ liệu.
 - » Mẫu kết quả khai phá được là những mẫu thể hiện tri thức nếu chúng dễ hiểu, hợp lệ với một mức độ chắc chắn, hữu dụng, và mới đối với người dùng.
 - » Lượng lớn dữ liệu từ các cơ sở dữ liệu truyền thống/hiện đại, kho dữ liệu, hay từ các nguồn thông tin khác (spatial, time series, text, multimedia, web, ...).

- » Các tác vụ khai phá dữ liệu bao gồm khai phá mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa dữ liệu), khai phá luật kết hợp/tương quan, phân lớp, dự đoán, gom cụm, phân tích xu hướng, phân tích độ lệch và phân tử biên, phân tích độ tương tự, ...
 - » Năm thành tố cơ bản để đặc tả một tác vụ khai phá dữ liệu: dữ liệu cụ thể sẽ được khai phá, loại tri thức sẽ đạt được, tri thức nền, các độ đo, và các kỹ thuật biểu diễn/trực quan hóa tri thức.
 - » Bốn thành phần cơ bản của một giải thuật khai phá dữ liệu: cấu trúc mẫu hay mô hình, hàm tỉ số, phương pháp tìm kiếm và tối ưu hóa, chiến lược quản lý dữ liệu.

- » Các tác vụ khai phá dữ liệu bao gồm khai phá mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa dữ liệu), khai phá luật kết hợp/tương quan, phân lớp, dự đoán, gom cụm, phân tích xu hướng, phân tích độ lệch và phân tử biên, phân tích độ tương tự, ...
 - » Năm thành tố cơ bản để đặc tả một tác vụ khai phá dữ liệu: dữ liệu cụ thể sẽ được khai phá, loại tri thức sẽ đạt được, tri thức nền, các độ đo, và các kỹ thuật biểu diễn/trực quan hóa tri thức.
 - » Bốn thành phần cơ bản của một giải thuật khai phá dữ liệu: cấu trúc mẫu hay mô hình, hàm tỉ số, phương pháp tìm kiếm và tối ưu hóa, chiến lược quản lý dữ liệu.

- » Khai phá dữ liệu được xem như là một phần của quá trình khám phá tri thức.
- » Quá trình khám phá tri thức là một chuỗi lặp gồm các bước: làm sạch dữ liệu, tích hợp dữ liệu, chọn lựa dữ liệu, biến đổi dữ liệu, khai phá dữ liệu, đánh giá mẫu, và biểu diễn tri thức.
- » Nhiều lĩnh vực khác nhau có liên quan với khai phá dữ liệu: công nghệ cơ sở dữ liệu, lý thuyết thống kê, học máy, khoa học thông tin, trực quan hóa, ...
- » Các vấn đề liên quan: phương pháp luận khai phá dữ liệu, vấn đề tương tác người dùng, khả năng co giãn dữ liệu và hiệu suất, vấn đề xử lý lượng lớn các kiểu dữ liệu khác nhau, vấn đề khai thác các ứng dụng khai phá dữ liệu cũng như sự ảnh hưởng xã hội của chúng.

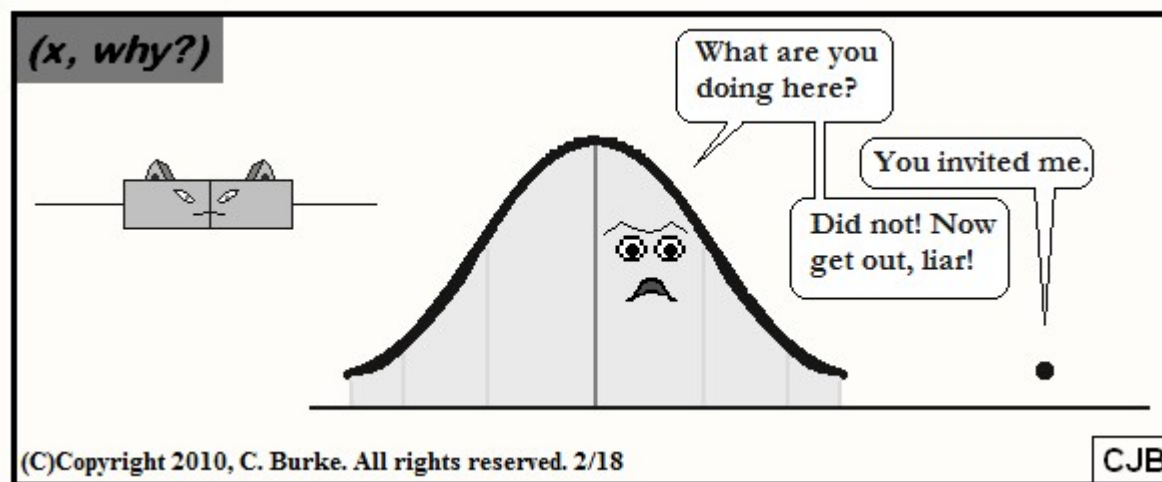


Anomaly Detection

Phát hiện bất thường

Phát hiện bất thường

25



Phát hiện bất thường

26

- » Về cơ bản, kỹ thuật khai phá dữ liệu (Data Mining) này dùng để nhấn mạnh vào việc quan sát các mục dữ liệu trong bộ dữ liệu để tìm ra các tập dữ liệu không khớp với mẫu dự kiến. Bất thường ở đây có thể đề cập đến độ lệch, sự khác thường, các nhiễu và ngoại lệ.
- » Sự bất thường được xem là khá quan trọng vì nó có thể cung cấp một số thông tin cần thiết. Nó có thể là một dữ liệu khác biệt so với mức trung bình chung trong một tập dữ liệu. Điều này chỉ ra rằng một cái gì đó khác thường đã xảy ra và các nhà phân tích dữ liệu cần chú ý.
- » Kỹ thuật này có thể được sử dụng trong nhiều lĩnh vực khác nhau. Chẳng hạn như phát hiện xâm nhập hay theo dõi sức khỏe.

Làm thế nào để phát hiện các điểm ngoại lai?

27

- Phát hiện giao dịch gian lận:
Có thể gọi $x(i)$ là tập các feature biểu diễn các activities của người dùng (số lần click chuột, số lần gửi request tới server, thời gian giữa các request...) Từ đó hệ thống có thể phát hiện người dùng có các hành động bất thường và đưa ra cảnh báo với người quản trị hệ thống
- Kiểm tra chất lượng trong dây chuyền sản xuất
- Giám sát các máy tính trong trung tâm dữ liệu
Các thuộc tính như bộ nhớ sử dụng, số lần truy cập nhớ, CPU load, network traffic có thể được sử dụng trong hoàn cảnh này.



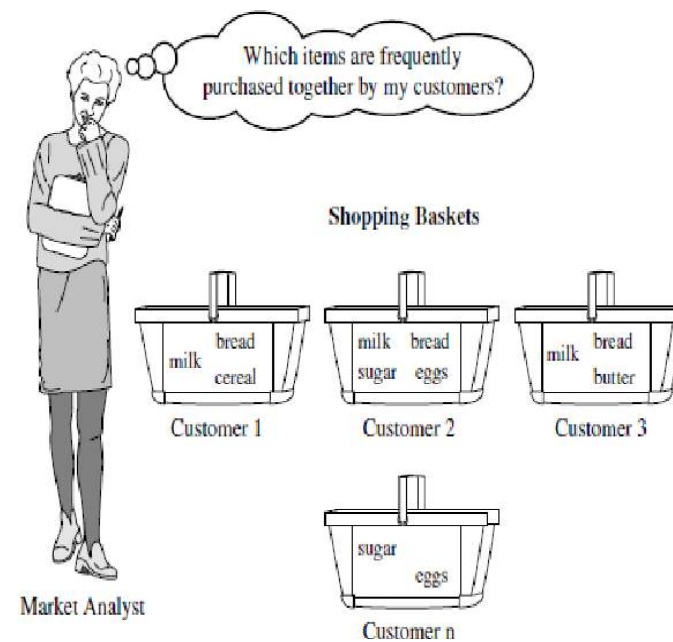
Association Rule Learning

Khai phá luật kết hợp

Bài toán phân tích

29

- Phân tích việc mua hàng của khách hàng bằng cách tìm ra những “mối kết hợp” giữa những mặt hàng mà khách đã mua



Bài toán phân tích

30

Những mặt hàng nào thường được khách hàng mua cùng nhau trong cùng 1 lần mua hàng?

- Thiết kế gian hàng.
- Lên kế hoạch bán giảm giá cho mặt hàng/nhóm mặt hàng.
- Lên kế hoạch tiếp thị/các chiến lược quảng cáo.
- .v.v.

Khai phá luật kết hợp

31

- Kỹ thuật Association Rule Learning trong khai phá dữ liệu được sử dụng để xác định mối quan hệ giữa các biến khác nhau trong cơ sở dữ liệu.
- Ngoài ra, nó còn được sử dụng để “giải nén” các mẫu ẩn trong dữ liệu. Association Rule rất hữu ích để kiểm tra, dự đoán hành vi và thường được áp dụng trong ngành bán lẻ.
- Các doanh nghiệp sử dụng kỹ thuật này để xác định hành vi mua sắm, phân tích dữ liệu trong giỏ hàng của khách hàng tiềm năng. Trong lĩnh vực Công nghệ Thông tin, các lập trình viên sử dụng kỹ thuật này để xây dựng các chương trình Machine Learning.

Khai phá luật kết hợp

32

- » Cho $I = \{I_1, I_2, \dots, I_n\}$ là một tập các mục (mặt hàng, .v.v.).
- » Cho D là một tập các giao dịch mà mỗi giao dịch T là một tập các mục, $T \subseteq I$.
- » Mỗi giao dịch có một mã định danh riêng gọi là TID .
- » Cho A là một tập các mục (mặt hàng). Một giao dịch T được gọi là chứa A khi và chỉ khi $A \subseteq T$.
- » Một **luật kết hợp** được diễn đạt dưới hình thức $A \Rightarrow B$, với $A \subset I, B \subset I$, và $A \cap B = \emptyset$

Khai phá luật kết hợp

33

- » Luật $A \Rightarrow B$ trong tập giao dịch D có **độ hỗ trợ - support** – s , với s là phần trăm số giao dịch trong D chứa $A \cup B$, (cả A và B).

$$\text{support}(A \Rightarrow B) = P(A \cup B)$$

- » Luật $A \Rightarrow B$ trong tập giao dịch D có **độ tin cậy - confidence** – c , với c là phần trăm giao dịch trong D có chứa A thì cũng chứa B .

$$\text{confidence}(A \Rightarrow B) = P(B|A)$$

Khai phá luật kết hợp

34

Association Rule chia làm 2 phần cần tìm

- “NẾU”: điều kiện tiên đề, nguyên nhân, cái đến trước
- “THÌ”: kết quả, cái đến sau

3 thông số để đánh giá:

- Support : Giá trị Support của 1 sản phẩm hoặc 1 tổ hợp sản phẩm là tỷ lệ giữa số giao dịch chứa sản phẩm hoặc tổ hợp sản phẩm đó so với tổng số giao dịch
- Confidence: Xác suất có điều kiện thể hiện khả năng khách hàng mua sản phẩm A cũng sẽ mua sản phẩm B
- Lift: Nếu khách hàng mua sản phẩm A thì khả năng họ mua sản phẩm B sẽ tăng lên bao nhiêu %

Khai phá luật kết hợp – Một số độ đo phổ biến

35

Những luật kết hợp thỏa mãn độ hỗ trợ tối thiểu (min_sup) và độ tin cậy tối thiểu (min_conf) được gọi là các **luật mạnh**.

Khai phá luật kết hợp – Tập mục thường xuyên

36

Một tập mục chứa k mục được gọi là một k -itemset.

Ví dụ:

» tập {computer, antivirus_software} là một 2-itemset

Tần suất xuất hiện của một tập mục là số giao dịch chứa tập mục đó (hay còn được gọi là *frequency*, *support count*, *count* của tập mục)

Nếu độ hỗ trợ (support) của một tập mục I lớn hơn hoặc bằng độ hỗ trợ tối thiểu (min_sup) thì I được gọi là một *tập mục thường xuyên*.

Khai phá luật kết hợp

37

VD: Cửa hàng có 100.000 giao dịch, trong đó:

2.500 giao dịch có sản phẩm dầu xả (2.5%)

10.000 giao dịch có sản phẩm dầu gội đầu (10%)

1.000 giao dịch có cả 2 sản phẩm trên

Giả sử luật kết hợp được tìm thấy “NẾU khách hàng mua dầu xả, THÌ mua dầu gội”

Support: $(1.000/100.000)=1\%$

Confidence: $(1.000/2.500)=40\%$

Lift tăng từ 10% lên 40%

Khai phá luật kết hợp

38

Transaction	Item purchase	Transaction	Asparagus	Beans	Broccoli	Corn	Green Peppers	Squash	Tomatoes
1	Broccoli, green peppers, corn	1	0	0	1	1	1	0	0
2	Asparagus, squash, corn	2	1	0	0	1	0	1	0
3	Corn, tomatoes, beans, squash	3	0	1	0	1	0	1	1
4	Green peppers, corn, tomatoes, beans	4	0	1	0	1	1	0	1
5	Beans, asparagus, broccoli	5	1	1	1	0	0	0	0
6	Squash, asparagus, beans, tomatoes	6	1	1	0	0	0	1	1
7	Tomatoes, corn	7	0	0	0	1	0	0	1
8	Broccoli, tomatoes, green peppers	8	0	0	1	0	1	0	1
9	Squash, asparagus, beans	9	1	1	0	0	0	1	0
10	Beans, corn	10	0	1	0	1	0	0	0
11	Green peppers, broccoli, beans, squash	11	0	1	1	0	1	1	0
12	Asparagus, beans, squash	12	1	1	0	0	0	1	0
13	Squash, corn, asparagus, beans	13	1	1	0	1	0	1	0
14	Corn, green peppers, tomatoes, beans, broccoli	14	0	1	1	1	1	0	1

Liệt kê theo từng giao dịch

Kiểu biến flag ứng với từng giao dịch

0: không chứa sản phẩm

1: có chứa sản phẩm

Khai phá luật kết hợp

39

Nhìn chung, khai phá luật kết hợp là một quá trình 2 bước.

- ***Tìm tất cả các tập mục thường gặp (thường xuyên)***: theo định nghĩa thì mỗi itemset được gọi là tập mục thường xuyên nếu độ hỗ trợ của nó lớn hơn hoặc bằng *min_sup*.
- ***Tạo các luật kết hợp mạnh từ các tập mục thường xuyên***: theo định nghĩa thì những luật kết hợp mạnh phải có độ hỗ trợ và độ tin cậy lớn hơn *min_sup* và *min_conf* tương ứng.

Các ứng dụng phổ biến

40

Market Basket Analysis: đây là ứng dụng chính, vì nó cho thấy lợi ích kinh doanh cao nhất trong các thuật toán Data mining được sử dụng, hầu như các nhà bán lẻ hay TMĐT đều biết đến.

Các ứng dụng phổ biến

41

Đối với website TMĐT:

- Hỗ trợ xây dựng hệ thống khuyến nghị
- Marketing target nhiều khách hàng chính xác hơn

Đối với cửa hàng truyền thống:

- Tối ưu việc trưng bày sản phẩm
- Đưa ra các khuyến mãi mua theo gói combo cho khách hàng
- Tăng cường hoạt động cross-selling, các nhân viên bán hàng có thể dễ dàng tư vấn thêm các sản phẩm liên quan cho khách

Các ứng dụng phổ biến

42

Chuẩn đoán bệnh trong y tế:

Khai phá luật kết hợp trong lĩnh vực y học mục đích hỗ trợ bác sĩ biết được xác suất xảy ra bệnh A bất kỳ khi liên kết hàng loạt các yếu tố, triệu chứng B,C. Ngoài ra còn cho phép phát hiện ra các triệu chứng mới, bệnh mới.



PHÂN TÍCH PHÂN LOẠI

Classification Analysis

Định nghĩa

44

Phân tích phân lớp là một hình thức phân tích dữ liệu phổ biến được dùng để tạo lập các mô hình mô tả các phân lớp dữ liệu quan trọng. Những phân tích như vậy giúp chúng ta hiểu sâu hơn về dữ liệu ở một tầm mức lớn hơn, bao quát hơn. Những mô hình phân lớp (classification). Đặc trưng của mô hình phân lớp là học dựa trên tập dữ liệu mẫu có sẵn. Những dữ liệu thô đầu vào (như: văn bản, hình ảnh, âm thanh, .v.v.) được chuyển đổi thành các vector. Số chiều của các vector này tùy thuộc vào từng trường hợp cụ thể. Đặc biệt trong các chiều này, có một chiều chứa nhãn phân lớp có sẵn tương ứng với từng trường hợp cụ thể.

Định nghĩa

45

Đặc trưng của mô hình phân lớp là học dựa trên tập dữ liệu mẫu có sẵn. Những dữ liệu thô đầu vào (như: văn bản, hình ảnh, âm thanh, .v.v.) được chuyển đổi thành các vector. Số chiều của các vector này tùy thuộc vào từng trường hợp cụ thể. Đặc biệt trong các chiều này, có một chiều chứa nhãn phân lớp có sẵn tương ứng với từng trường hợp cụ thể.

Dựa trên tập dữ liệu vào dạng vector này, các thuật toán đặc thù sẽ được triển khai để tiến hành thiết lập một ánh xạ $f(x)=c, c \in C$. Trong đó C là tập các nhãn và x là vector với các chiều. $f(x)$ chính là mô hình thu được thông qua quá trình học.

Các phương pháp phân lớp

46

Chuẩn bị dữ liệu cho việc huấn luyện mô hình phân lớp nhằm phục vụ hai mục đích chính:

1. Xây dựng mô hình phân lớp: tập dữ liệu dùng cho mục đích này được gọi là tập dữ liệu huấn luyện (*training dataset*).
2. Kiểm thử mô hình phân lớp: tập dữ liệu dùng cho mục đích này được gọi là tập dữ liệu kiểm thử (*test set/validation set*). Tập dữ liệu này được sử dụng để kiểm tra độ chính xác về khả năng dự đoán của mô hình phân lớp.

Chuẩn bị dữ liệu

47

Làm sạch dữ liệu

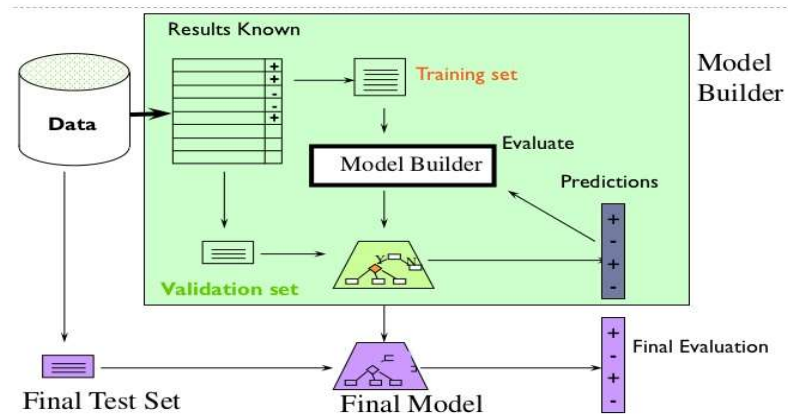
- Nhiều
- Thiếu giá trị

Phân tích liên quan (chọn đặc trưng)

- Các thuộc tính không liên quan
- Các thuộc tính dư thừa

Biến đổi dữ liệu

Classification: Train, Validation, Test split



Các phương pháp phân lớp

48

- Các phương pháp cây quyết định
 - » Decision Tree based Methods
- Các phương pháp dựa trên luật
 - » Rule-based Methods
- Các phương pháp Bayes «ngây thơ» và mạng tin cậy Bayes
 - » Naïve Bayes and Bayesian Belief Networks
- Các phương pháp máy vector hỗ trợ
 - » Support Vector Machines
- Lập luận dựa trên ghi nhớ
 - » Memory based reasoning
- Các phương pháp mạng nơron
 - » Neural Networks



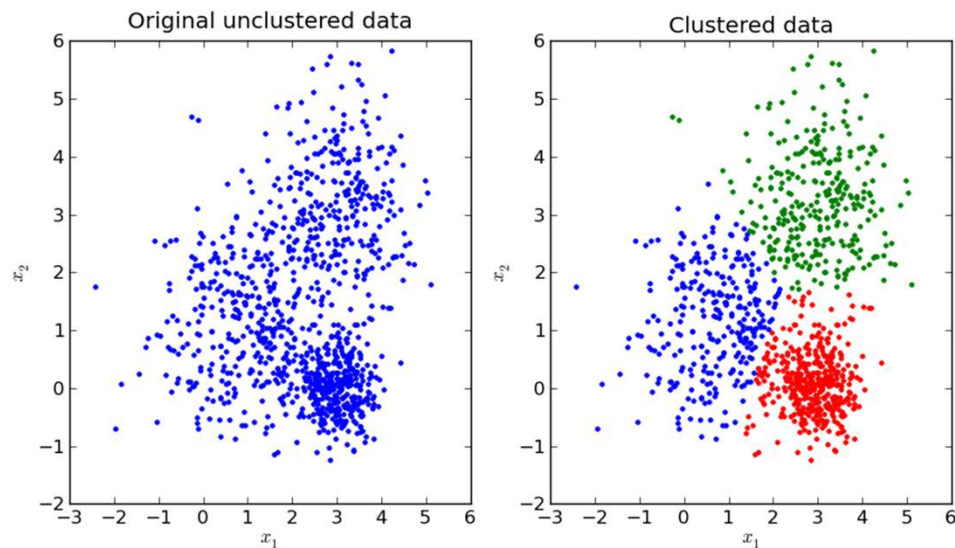
PHÂN TÍCH PHÂN CỤM

Cluster Analysis

Định nghĩa

50

Phương pháp phân tích cụm là phương pháp phân tích qua đó tập dữ liệu sẽ được phân thành nhiều cụm/nhóm khác nhau, trong mỗi cụm/nhóm các điểm dữ liệu hay các quan sát sẽ giống nhau, và giữa các cụm/nhóm có sự khác biệt.



Thế nào là PP phân cụm tốt?

51

- » Có độ tương tự cao trong cùng cụm (intra-class)
- » Có độ tương tự thấp giữa các cụm (inter-class)
- » Khả năng phát hiện mẫu ẩn (hidden patterns)
- » Có khả năng làm việc hiệu quả với mẫu lớn (scalability)
- » Khả năng làm việc với nhiều loại dữ liệu khác nhau

Các phương pháp phân cụm

52

- » Thuật toán phân hoạch (Partitioning algorithms)
Phân hoạch cơ sở dữ liệu D có n đối tượng thành k cụm:
 - Mỗi cụm có ít nhất 1 đối tượng
 - Mỗi đối tượng thuộc về 1 cụm duy nhất
 - K là số 1 cụm cho trước

Các phương pháp phân cụm

53

» Thuật toán phân cấp (Hierarchy algorithms)

– Gộp:

- Xuất phát mỗi đối tượng và tạo một cụm chứa nó.
- Nếu 2 cụm gần nhau thì gộp thành 1 cụm
- Lặp lại bước 2 cho đến khi còn 1 cụm duy nhất là toàn bộ không gian

– Tách:

- Xuất phát từ 1 cụm duy nhất là toàn bộ không gian
- Chọn cụm có độ phân biệt cao nhất (ma trận phân biệt có phần tử lớn nhất hoặc giá trị trung bình lớn nhất) để tách đôi
- Lặp lại bước 2 cho đến khi mỗi đối tượng thuộc 1 cụm hoặc đạt điều kiện dừng (đủ số cụm hoặc khoảng cách giữa các cụm đủ nhỏ)

Các phương pháp phân cụm

54

- » Phương pháp dựa trên mật độ (Densitybased methods)
- » Phương pháp dựa trên lưới (Grid-based methods)
- » Phương pháp dựa trên mô hình (Modelbased)

Các ứng dụng phổ biến

55

Trong kinh doanh:

Phân nhóm khách hàng trong tập dữ liệu theo các đặc tính khác nhau kết hợp với những hiểu biết, kinh nghiệm về các yếu tố thu hút khách hàng. Từ đó có thể tinh chỉnh các giải pháp bán hàng, marketing phù hợp



Các ứng dụng phổ biến

56

Những dữ liệu về khách hàng có thể sử dụng phân cụm:

- » Browsing website của khách
- » Nhân khẩu học
- » Hành vi của khách hàng trên các kênh TMĐT, mạng xã hội
- » Lịch sử giao dịch
- » Các cuộc đối thoại, hội thoại qua tổng đài
- » Mức độ thường xuyên giao dịch

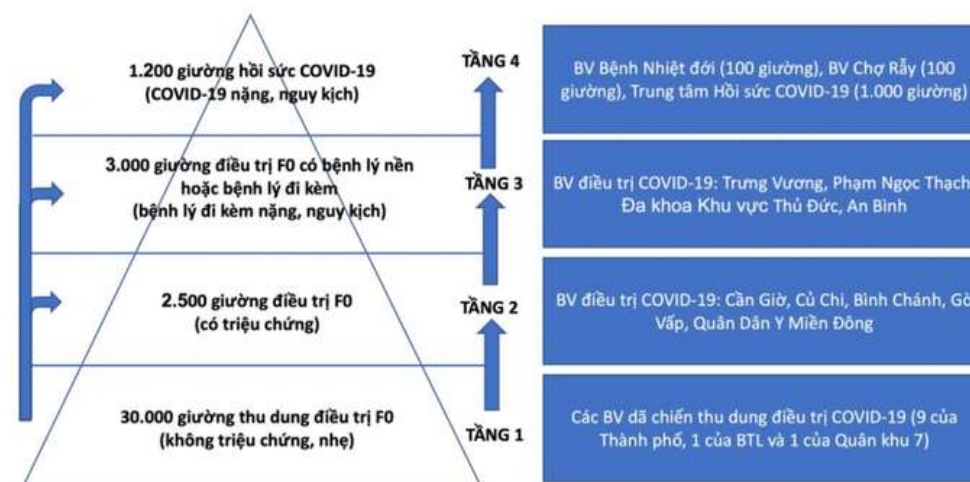


Các ứng dụng phổ biến

57

Trong y tế, sức khỏe: được ứng dụng trong tâm lý học, hỗ trợ việc cải thiện và duy trì sức khỏe, nâng cao hệ thống chăm sóc sức khỏe, và ngăn ngừa các loại bệnh.

MÔ HÌNH THÁP 4 TẦNG TRONG ĐIỀU TRỊ COVID-19



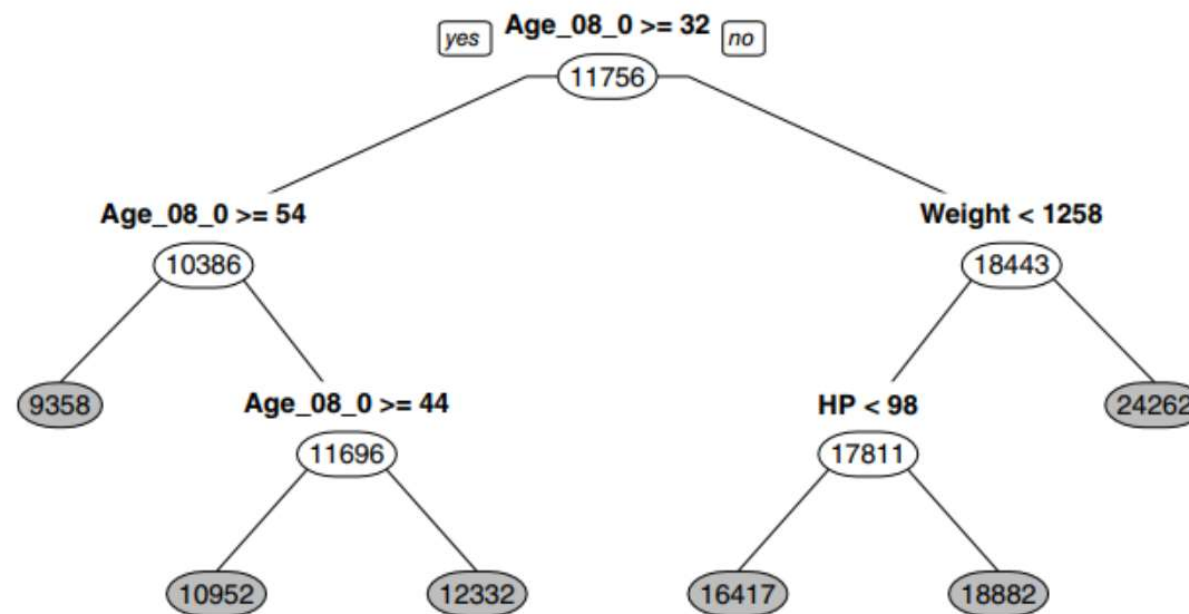


PHÂN TÍCH HỒI QUY

Regression Analysis

Định nghĩa

59



Định nghĩa

60

- » **Phân tích hồi qui** là kỹ thuật thống kê dùng để ước lượng phương trình phù hợp nhất với các tập hợp kết quả quan sát của biến phụ thuộc và biến độc lập
- » **Phân tích hồi qui** cho phép đạt được kết quả ước lượng tốt nhất về mối quan hệ chân thực giữa các biến số. Từ phương trình ước lượng được này, người ta có thể dự báo về biến phụ thuộc (chưa biết) dựa vào giá trị cho trước của biến độc lập (đã biết)..

Phân loại

61

- » Hồi qui tuyến tính (linear) và phi tuyến (nonlinear)
- » Hồi qui đơn biến (single) và đa biến (multiple)
- » Cây hồi quy (regression trees)

Ứng dụng

62

- ✓ Quá trình khai phá dữ liệu
 - » Giai đoạn tiền xử lý dữ liệu
 - » Giai đoạn khai phá dữ liệu
 - Khai phá dữ liệu có tính mô tả
 - Khai phá dữ liệu có tính dự báo
- ✓ Trong Kinh doanh:
 - Đưa ra các dự báo trong tương lai
 - Tối ưu hóa quá trình hoạt động
 - Hỗ trợ ra quyết định
 - Hỗ trợ khai phá các thông tin hữu ích và có giá trị