Khai phá luật kết hợp

Tìm tần số mẫu, mối kết hợp, sự tương quan hay các cấu trúc nhân quả giữa các tập đối tượng trong các cơ sở dữ liệu giao tác, cơ sở dữ liệu quan hệ, và những kho thông tin khác

Yêu cầu:

- **Tính hiểu được: dễ hiểu**
- **Tính sử dụng được: cung cấp thông tin thiết thực**
- **Tính hiệu quả:**

Ứng dụng:

- **Phân tích dữ liệu giỏ hàng, thương mại,..**
- **Thiết kế catalogue, website, đồ họa,...**
- **Sinh học, sữa chữa,...**

Khai phá luật kết hợp

Định dạng thể hiện đặc trưng cho các luật kết hợp:

- **Khăn → Bia [0.5%,60%]**
- **Mua:Khăn → Mua:Bia [0.5%,60%]**
- **NẾU mua khăn THÌ mua bia trong 60% trường hợp. Khăn và bia được mua chung trong 0.5% dòng dữ liệu**

Các hình thức biểu diễn khác:

- **Mua(x,"khăn") → Mua(x,"bia") [0.5%,60%]**
- **Khoa(x, "CNTT") ^ Học(x,"DB") → Điểm(x,"A") [1%,75%]**

Các hướng tiếp cận luật kết hợp

Luật kết hợp nhị phân (Binary Association Rule):

Các thuộc tính chỉ được quan tâm là có hay không xuất hiện trong giao tác của cơ sở dữ liệu (không quan tâm về "mức độ" xuất hiện)

Ví dụ:

- **Việc gọi 10 cuộc điện thoại và 1 cuộc được xem là giống nhau (có cuộc gọi hay không – Có hay Không?)**
- **NẾU "gọi liên tỉnh='yes' AND gọi di động="yes"**
  THÌ gọi quốc tế='yes' AND gọi dịch vụ 108 = 'yes'
  với độ hỗ trợ 20% và độ tin cậy 80%"

Các hướng tiếp cận luật kết hợp

Luật kết hợp có thuộc tính số và thuộc tính hạng mục (Quantitative And Categorial Association Rule)

Các thuộc tính của các cơ sở dữ liệu có kiểu đa dạng (nhị phân – binary, số – quantitative, hạng mục – categorial, ...)

→ Rời rạc hoá nhằm chuyển dạng luật này về dạng nhị phân

Ví dụ:

NẾU phương thức gọi=’Tự động’

AND giờ gọi ∈ ‘23:00...23:59’

AND Thời gian đàm thoại ∈ ‘20.. 30 phút’

THÌ gọi liên tỉnh =’có’ ,

với độ hỗ trợ là 23. 53% , và độ tin cậy là 80%”.

Các hướng tiếp cận luật kết hợp

Luật kết nhiều mức (Multi-level Association Rule)

Dạng luật đầu là dạng luật tổng quát hoá của dạng luật sau và tổng quát theo nhiều mức khác nhau

Ví dụ:

Luật có dạng:

NẾU mua máy tính PC

THÌ mua hệ điều hành

AND mua phần mềm tiện ích văn phòng

thay vì chỉ những luật quá cụ thể:

NẾU mua máy tính iBM PC

THÌ mua hệ điều hành Microsoft Windows

AND mua phần mềm Microsoft Office

Các hướng tiếp cận luật kết hợp

Luật kết hợp mờ (Fuzzy Association Rule)

Trong quá trình rời rạc hoá các thuộc tính số, luật kết hợp mờ nhằm khắc phục các hạn chế và chuyển luật kết hợp về một dạng tự nhiên hơn, gần gũi hơn với người sử dụng

Ví dụ:

NẾU thuê bao tư nhân = 'yes'

AND thời gian đàm thoại lớn ***(Thuộc tính được mờ hóa)***

AND cước nội tỉnh = 'yes'

THÌ cước không hợp lệ = 'yes'

với độ hỗ trợ 4% và độ tin cậy 85%".

Khai phá luật kết hợp

Phân tích định dạng luật kết hợp:

NẾU mua khăn THÌ mua bia trong 60% trường hợp. Khăn và bia được mua chung trong 0.5% dòng dữ liệu

Khăn → Bia [0.5%,60%]

1. **Tiền đề: Khăn (vế trái)**
2. **Mệnh đề kết quả: Bia (vế phải, đầu)**
3. **Support: 0.5% - tần số (hay độ hỗ trợ, độ phổ biến) – trong bao nhiêu % dữ liệu thì những điều ở vế trái và vế phải cùng xảy ra?**
4. **Confidence: 60% - độ mạnh (hay xác suất điều kiện, độ tin cậy, độ gắn kết) – nếu vế trái xảy ra thì có bao nhiêu khả năng vế phải xảy ra?**

Khai phá luật kết hợp

Độ ủng hộ: Biểu thị tần số luật có trong các giao tác

$$support(A \Rightarrow B\ [\ s,\ c\ ]) = p(A \cup B) = \underline{support\ (\{A,B\})}$$

Độ tin cậy: biểu thị số phần trăm giao tác có chứa luôn B trong các giao tác có chứa A

$$confidence(A \Rightarrow B\ [\ s,\ c\ ]) = p(B|A) = p(A \cup B) / p(A) = \underline{support(\{A,B\}) / support(\{A\})}$$

Khai phá luật kết hợp

Độ ủng hộ tối thiểu (min support):

- Cao: → ít tập phần tử (itemset) phổ biến
  → ít luật hợp lệ rất thường xuất hiện
- Thấp: → nhiều luật hợp lệ hiếm xuất hiện

Độ tin cậy tối thiểu (min confidence):

- Cao: → ít luật nhưng tất cả "gần như đúng"
- Thấp: → nhiều luật, phần lớn rất "không chắc chắn"

Giá trị tiêu biểu:

minsupport: 2-10%, minconfidence: 70-90%

Khai phá luật kết hợp

item và itemsets:

i = { i_1, i_2, …, i_n } là tập bao gồm n mục (item – còn gọi là thuộc tính – attribute). X ⊆ i được gọi là tập mục (itemset).

Giao tác:

T = { t_1, t_2, …, t_m} là tập gồm m giao tác (Transaction – còn gọi là bản ghi –record). Mỗi giao tác được định danh bởi TiD (Transaction identification).

Tập phần tử phổ biến:

Tập các phần tử có độ ủng hộ (support) ≥ độ ủng hộ tối thiểu (minsupport)