KHAI PHÁ DỮ LIỆU

Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giáo viên: TS.Nguyễn Khắc Cường

CHỦ ĐỀ 6

LUẬT KẾT HỢP
(Association rule)

Luật kết hợp

- Giới thiệu bài toán
  - Trong thực tế:
    - Bán hàng trực tiếp:
      - Ghi nhận:
        - thông tin các loại hàng hóa khách hàng đã mua
      - Cần:
        - tìm ra các loại hàng hóa nào có nhiều khả năng được mua cùng nhau nhất
      - Sử dụng:
        - tận dụng thông tin đó để điều chỉnh chiến lược kinh doanh, tiếp thị, bán hàng.
        - Ví dụ: xếp các loại hàng thường được mua cùng nhau ở gần nhau

Luật kết hợp

- Giới thiệu bài toán
  - Trong thực tế:
    - Bán hàng online:
      - Tương tự bán hàng trực tiếp
      - Sử dụng:
        - Khi khách hàng chọn mua một món hàng, trang bán hàng chọn các món hàng liên quan (có khả năng mua nhiều) đến món hàng khách đã chọn để mời khách hàng mua thêm
    - Y tế:
      - Ghi nhận thông tin:
        - các triệu chứng, các loại bệnh, các loại thuốc đã dùng
      - Cần tìm ra:
        - các triệu chứng có khả năng cao xuất hiện cùng nhau
        - các loại bệnh xuất hiện cùng nhau
        - các loại thuốc dùng cùng nhau

Luật kết hợp

- Giới thiệu bài toán
  - Trong thực tế:
    - Y tế:
      - Sử dụng các thông tin tìm được để :
        - Điều chỉnh các chẩn đoán bệnh được chính xác hơn
        - Điều chỉnh các đơn thuốc hiệu quả hơn
    - . . .
  - Trong data mining:
    - Luật kết hợp được dùng để tìm trong dataset các tập dữ liệu có
      - Có mối tương quan với nhau
      - Có sự xuất hiện cùng nhau
    - Các giải thuật giúp tìm ra các tập dữ liệu như trên được gọi là các giải thuật khai phá luật kết hợp

Luật kết hợp

- Các khái niệm:
  - Dataset: lưu giữ một số lượng lớn dữ liệu (features, samples) đã xảy ra ở các giao dịch trong quá khứ
  - Item: là một feature trong dataset
  - Itemset: tập hợp các features xuất hiện cùng nhau trong các giao dịch
  - i-itemset: là itemset có i feature
  - Support: là khả năng (%) xuất hiện một itemset trong các giao dịch
- Phát biểu bài toán:
  - Gọi $I = \{i_1, i_2, \ldots, i_m\}$ là tập hợp các item (feature)

Luật kết hợp

- Phát biểu bài toán:
  - DB: lưu tập hợp các giao dịch
  - T: là một giao dịch, chứa một tập các items ($T \subseteq I.$)
  - TID: tên của từng giao dịch, ví dụ

Database DB:

TID	Items
001	*A C D*
002	*B C E*
003	*A B C E*
004	*B E*

  - X: một tập các items, một giao dịch T được gọi là chứa X nếu $X \subseteq T.$

Luật kết hợp

- Phát biểu bài toán:
  - $X \Rightarrow Y$.: là một luật kết hợp, trong đó $X \subseteq I, Y \subseteq I, X \cap Y = \emptyset$.
  - Confidence c: là khi DB chứa các items X thì có c% khả năng DB cũng chứa các items Y
  - Support s: là có khả năng s% trong DB có chứa $X \cup Y$.
  - Bài toán khai phá luật kết hợp gồm:
    - Xác định các itemset
    - Sử dụng các itemset để tìm ra các luật kết hợp có confidence c cao hơn một giá trị xác định
- Giải thuật Apriori:
  - Là giải thuật giúp tìm ra các itemset thường xuất hiện trong dataset + xác định luật kết hợp mạnh

Luật kết hợp

- Giải thuật Apriori:
  - Lặp nhiều lần, mỗi bước lặp thứ i
    - tìm các i-itemset thường xuất hiện nhất → thực hiện qua 2 bước:
      - Candidate generation
      - Candidate counting and selection
- Minh họa cụ thể:
  - Tìm tất cả các 1-itemset
  - Tính support của từng 1-item
  - Chọn những 1-itemset nào có support > pre-defined threshold → gọi là những 1-itemset xuất hiện thường xuyên

Luật kết hợp

- Minh họa cụ thể:
  - Xét DB

Database DB:

TID	Items
001	*A C D*
002	*B C E*
003	*A B C E*
004	*B E*

  - Các 1-itemset

1-Itemsets C_1
{*A*}
{*C*}
{*D*}
{*B*}
{*E*}

Luật kết hợp

- Minh họa cụ thể:
  - Tính support của từng 1-itemset

1-Itemsets	Count	s[%]
{A}	2	50
{C}	3	75
{D}	1	25
{B}	3	75
{E}	3	75

  - Giả sử chọn threshold = 50% → chọn ra các 1-itemset thường xuất hiện là

Large 1-itemsets $\mathbf{L_1}$	Count	s[%]
{A}	2	50
{C}	3	75
{B}	3	75
{E}	3	75

Luật kết hợp

- Minh họa cụ thể:
  - Tạo các 2-itemset:

$$L_k * L_k = \{X \cup Y \text{ where } X, Y \in L_k, |X \cap Y| = k-1\}$$

  - Số lượng các 2-itemset

$$|L_1| \cdot (|L_1| - 1)/2 = 4 \cdot 3/2 = 6$$

2-Itemsets $\mathbf{C_2}$
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

Luật kết hợp

- Minh họa cụ thể:
  - Tính support của các 2-itemset:

2-Itemsets	Count	s[%]
$\{A, B\}$	1	25
$\{A, C\}$	2	50
$\{A, E\}$	1	25
$\{B, C\}$	2	50
$\{B, E\}$	3	75
$\{C, E\}$	2	50

  - Chọn các 2-itemset (threshold = 50%)

Large 2-Itemsets $\mathbf{L_2}$	Count	s[%]
$\{A, C\}$	2	50
$\{B, C\}$	2	50
$\{B, E\}$	3	75
$\{C, E\}$	2	50

Luật kết hợp

- Minh họa cụ thể:
  - Tạo các 3-itemset (threshold = 50%)

$$L_k * L_k = \{X \cup Y \text{ where } X, Y \in L_k, |X \cap Y| = k-1\}$$

→ Gồm có: {A,B,C}, {A,B,E}, {B,C,E}, …

- Do có các tập con 2-itemset của {A,B,C}, {A,B,E} không phải là large 2-itemset → loại
- Chỉ có {B,C,E} là có tất cả các tập con 2-itemset đều là large itemset → chọn

3-Itemsets C_3
{B, C, E}

Luật kết hợp

- Minh họa cụ thể:
  - Tính support

3-Itemsets	Count	s[%]
$\{B, C, E\}$	2	50

  - Chọn (threshold = 50%)

Large 3-Itemsets $\mathbf{L_3}$	Count	s[%]
$\{B, C, E\}$	2	50

  - Do không tạo được 4-itemset → giải thuật dừng
  - Kết quả: tìm được các item thường xuất hiện là: {B,C,E}

→ Hoàn thành bước “Candidate generation”

Luật kết hợp

- Minh họa cụ thể:
  - Bước "Candidate counting and selection"
    - Xét từng luật kết hợp có thể có: {B} → C, {B} → E, {C} → E, {B,C} → E, …
    - Tính từng confidence của từng luật
    - Chọn luật nào > threshold làm luật mạnh
    - Ví dụ xét luật kết hợp {B,C} → E có là luật mạnh hay không?
      - Xét support của các tập {B,C} và {B,C,E} trong bảng L_2 và L_3

$$s(B,C)=2, \quad s(B,C,E)=2$$

      - Tính confidence của luật kết hợp {B,C} → E :

$$c(\{B,C\}\rightarrow E)=\frac{s(B,C,E)}{s(B,C)}=\frac{2}{2}=1\,(\text{or } 100\%)$$

    → luật {B,C} → E có confidence > bất kỳ threshold nào

    → luật {B,C} → E là một luật mạnh

Luật kết hợp

- Ứng dụng của luật kết hợp:
  - Dữ liệu dùng để tìm luật → đã xảy ra trong quá khứ
  - Xác định luật mạnh → để dự đoán nếu có sự kiện nào đó xảy ra thì khả năng xảy ra của một sự kiện liên quan là bao nhiêu

  → Dùng để hỗ trợ cho nhiều lĩnh vực trong thực tế

Q / A