KBS dựa trên xác suất

TS. NGUYỄN ĐÌNH HÓA

Cách thức xây dựng KBS

Thu nhận và biểu diễn tri thức

- Kỹ sư tri thức thu nhận tri thức từ các chuyên gia trong từng lĩnh vực.
- Biểu diễn các luật mô tả trong từng lĩnh vực cụ thể
- Nhiều khi không được biểu diễn tri thức tường minh theo luật, sự kiện hay các quan hệ
- Cần phát triển các hệ thống có khả năng học thông tin

Học là quá trình xác định các vấn đề chưa biết:

- Biểu diễn mối liên quan giữa giả thiết và kết luận
- Học từ ký hiệu: hình thức hóa, sửa chữa các luật tường minh, sự kiện, và các quan hệ
- Học từ dữ liệu: áp dụng cho những hệ thống được mô hình hóa bởi các tham số. Học bằng các kỹ thuật tối ưu các tham số dựa trên dữ liệu.

Học từ ký hiệu - Thuật toán hỗn loạn

Lý thuyết thông tin cho công thức tính độ hỗn loạn của tập dữ liệu sau khi phân lớp theo một thuộc tính A nào đó (chia tập dữ liệu thành k tập con dựa trên k giá trị của A)

$$\text{Độ } hỗn\ loạn\ trung\ bình = \sum_{b,c}\left(\frac{n_b}{n}\right)\left(-\frac{n_{bc}}{n_b}\log\frac{n_{bc}}{n_b}\right) = \sum_{b}\left(\frac{n_b}{n}\right)\left(\sum_{c}\left(-\frac{n_{bc}}{n_b}\log\frac{n_{bc}}{n_b}\right)\right)$$

Trong đó:

n là tổng số trường hợp xảy ra

$n_b$là tổng số trường hợp tương ứng với một giá trị thuộc tính A nào đó

n_{bc}là tổng số trường hợp tương ứng với một giá trị của A thuộc một lớp c nào đó

Cần lựa chọn thuộc tính có độ hỗn loạn trung bình nhỏ nhất để thực hiện phân lớp

ví dụ

Tìm quy luật chơi Tennis dựa trên thông tin từ bảng bên.

Tất cả giá trị của Outlook: 3

Tất cả các giá trị của Temperature: 3

Tất cả các giá trị của Humidity: 2

Tất cả các giá trị của windy: 2

Tất cả các nhãn Play: 2

TT	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Thuật toán hỗn loạn – ví dụ

Độ hỗn loạn theo từng thuộc tính:

$$E_{Outlook} = \frac{5}{14}\left[-\frac{2}{5}\,log_2\frac{2}{5} - \frac{3}{5}\,log_2\frac{3}{5}\right] + \frac{4}{14}\left[-\frac{4}{4}\,log_2\frac{4}{4} - \frac{0}{4}\,log_2\frac{0}{4}\right] + \frac{5}{14}\left[-\frac{3}{5}\,log_2\frac{3}{5} - \frac{2}{5}\,log_2\frac{2}{5}\right] = \mathbf{0,69}$$

$$E_{Temperature} = \frac{4}{14}\left[-\frac{2}{4}\,log_2\frac{2}{4} - \frac{2}{4}\,log_2\frac{2}{4}\right] + \frac{6}{14}\left[-\frac{4}{6}\,log_2\frac{4}{6} - \frac{2}{6}\,log_2\frac{2}{6}\right] + \frac{4}{14}\left[-\frac{3}{4}\,log_2\frac{3}{4} - \frac{1}{4}\,log_2\frac{1}{4}\right] = 0{,}91$$

$$E_{Humidity} = \frac{7}{14}\left[-\frac{2}{7}\,log_2\frac{2}{7} - \frac{5}{7}\,log_2\frac{5}{7}\right] + \frac{7}{14}\left[-\frac{6}{7}\,log_2\frac{6}{7} - \frac{1}{7}\,log_2\frac{1}{7}\right] = 0{,}73$$

$$E_{Windy} = \frac{6}{14}\left[-\frac{3}{6}\,log_2\frac{3}{6} - \frac{3}{6}\,log_2\frac{3}{6}\right] + \frac{8}{14}\left[-\frac{6}{8}\,log_2\frac{6}{8} - \frac{2}{8}\,log_2\frac{2}{8}\right] = 0{,}89$$

Chọn theo thuộc tính có độ hỗn loạn là nhỏ nhất (thuộc tính Outlook)

Xác định luật theo thuộc tính Outlook.

Luật 1: IF “Outlook” là “Overcast” THEN “Play” là “Yes”

Thuật toán hỗn loạn – ví dụ

Độ hỗn loạn theo tổ hợp 2 thuộc tính:

$$E_{(A_1 \text{là } Sunny) \cap A_2} = \frac{2}{5}\left[-\frac{0}{2}\,log_2\frac{0}{2} - \frac{2}{2}\,log_2\frac{2}{2}\right] + \frac{2}{5}\left[-\frac{1}{2}\,log_2\frac{1}{2} - \frac{1}{2}\,log_2\frac{1}{2}\right] + \frac{1}{5}\left[-\frac{1}{1}\,log_2\frac{1}{1} - \frac{0}{1}\,log_2\frac{0}{1}\right] = 0{,}4$$

$$E_{(A_1 \text{là } Sunny) \cap A_3} = \frac{3}{5}\left[-\frac{0}{3}\,log_2\frac{0}{3} - \frac{3}{3}\,log_2\frac{3}{3}\right] + \frac{2}{5}\left[-\frac{2}{2}\,log_2\frac{2}{2} - \frac{0}{2}\,log_2\frac{0}{2}\right] = \mathbf{0}$$

$$E_{(A_1 \text{là } Sunny) \cap A_4} = \frac{2}{5}\left[-\frac{1}{2}\,log_2\frac{1}{2} - \frac{1}{2}\,log_2\frac{1}{2}\right] + \frac{3}{5}\left[-\frac{1}{3}\,log_2\frac{1}{3} - \frac{2}{3}\,log_2\frac{2}{3}\right] = 0{,}95$$

Chọn tiêu chí có entropy nhỏ nhất: kết hợp thuộc tính Outlook=“Sunny” và Humidity.

Tạo ra các luật biểu diễn thông tin theo Outlook=“Sunny” và từng giá trị của Humidity:

Luật 2: IF “Outlook” là “Sunny” and “Humidity” là “High” THEN “Play” là “No”

Luật 3: IF “Outlook” là “Sunny” and “Humidity” là “Normal” THEN “Play” là “Yes”

Thuật toán hỗn loạn – ví dụ

Tương tự với một giá trị khác của Outlook:

$$E_{(A_1 \text{là } Rainy) \cap A_2} = \frac{0}{5}\left[-\frac{0}{0}\log_2\frac{0}{0} - \frac{0}{0}\log_2\frac{0}{0}\right] + \frac{3}{5}\left[-\frac{2}{3}\log_2\frac{2}{3}\ \frac{1}{3} - \log_2\frac{1}{3}\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] = 0{,}95$$

$$E_{(A_1 \text{là } Rainy) \cap A_3} = \frac{2}{5}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] + \frac{3}{5}\left[-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right] = 0{,}95$$

$$E_{(A_1 \text{là } Rainy) \cap A_4} = \frac{2}{5}\left[-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right] + \frac{3}{5}\left[-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right] = \mathbf{0}$$

Chọn theo tiêu chí entropy nhỏ nhất: Outlook kết hợp với Windy

Xây dựng thêm các luật mới:

Luật 4: IF “Outlook” là “Rainy” and “Windy” là “True” THEN “Play” là “No”

Luật 5: IF “Outlook” là “Rainy” and “Windy” là “False” THEN “Play” là “Yes”

Học từ dữ liệu - Thuật toán Bayes

Xây dựng các luật dựa trên xác suất có điều kiện.

Xác suất có điều kiện là xác suất xảy ra của một sự kiện ngẫu nhiên X khi biết sự kiện liên quan Y

Xác suất có điều kiện bị phụ thuộc vào các yếu tố:

- Xác suất tiên nghiệm: là xác suất xảy ra của riêng X, ký hiệu: P(X)
- Xác suất xảy ra của riêng Y, ký hiệu: P(Y)
- Xác suất xảy ra Y khi biết X, ký hiệu: P(Y|X)

Định lý Bayes:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Thuật toán Bayes – ví dụ

Xác định khả năng chơi tennis

TT	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Thuật toán Bayes – ví dụ

Trường hợp 1: chưa xuất hiện trong dữ liệu đã có

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

Trường hợp 2: đã xuất hiện trong dữ liệu đã có

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	?

Thuật toán Bayes – Ví dụ

Trường hợp 1:

Phân hoạch theo từng thuộc tính:

Outlook			Temp			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Wild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								

Tính toán dựa trên Bayes: tìm P(X|Y)

- X: khả năng chơi tennis: Yes/No
- Y: sự kiện xảy ra với tất cả các thuộc tính: $Y = (Y_1, Y_2, \ldots, Y_n)$

Outlook	Temp	Humidity	Windy
Sunny	**Cool**	**High**	**True**

Thuật toán Bayes – ví dụ

$$P(Yes|Y) = \frac{P(Outlook{=}Sunny|Yes).P(Temp{=}Cool|Yes).P(Humidity{=}High|Yes).P(Windy{=}True|Yes)}{P(Y)}$$

$$= \frac{(2/9\,.\,3/9\,.\,3/9\,.\,3/9)\,.\,9/14}{P(Y)} = \frac{0{,}007055}{P(Y)}$$

$$P(No|Y) = \frac{P(Outlook{=}Sunny|No).P(Temp{=}Cool|No).P(Humidity{=}High|No).P(Windy{=}True|No)}{P(Y)}$$

$$= \frac{(3/5\,.\,1/5\,.\,4/5\,.\,3/5)\,.\,5/14}{P(Y)} = \frac{0{,}027429}{P(Y)}$$

Kết luận: ước lượng xác suất dự báo cho mẫu tin *X* cho lớp “Play” là “Yes” nhỏ hơn ước lượng xác suất lớp “Play” là “No”,

Bayes đơn giản gán nhãn X cho lớp “Play” là “No”.

Thuật toán Bayes – Ví dụ

Trường hợp 2:

$$P(Yes|Y) = \frac{P(Outlook{=}Sunny|Yes).P(Temp{=}Hot|Yes).P(Humidity{=}High|Yes).P(Windy{=}False|Yes)}{P(Y)}$$

$$= \frac{(2/9\,.\,2/9\,.\,3/9\,.\,6/9)\,.\,9/14}{P(Y)} = \frac{0{,}007055}{P(Y)}$$

$$P(No|Y) = \frac{P(Outlook{=}Sunny|No).P(Temp{=}Hot|No).P(Humidity{=}High|No).P(Windy{=}False|No)}{P(Y)}$$

$$= \frac{(3/5\,.\,2/5\,.\,4/5\,.\,2/5)\,.\,5/14}{P(Y)} = \frac{0{,}027429}{P(Y)}$$

Kết luận: ước lượng xác suất dự báo cho mẫu tin *X* cho lớp “Play” là “Yes” nhỏ hơn ước lượng xác suất lớp “Play” là “No”,

Bayes đơn giản gán nhãn X cho lớp “Play” là “No”.

Thuật toán Bayes – bài tập

Hãy tìm khả năng chơi tennis trong điều kiện sau:

Outlook	Temp	Humidity	Windy	Play
Overcast	Mild	Normal	False	?