

Thuật giải ILA được đánh giá mạnh hơn hai thuật giải về phương pháp học quy nạp trước đây là ID3

### **Xác định dữ liệu**

1. Tập mẫu được liệt kê trong một bảng, với mỗi dòng tương ứng một mẫu, và mỗi cột thể hiện một thuộc tính trong mẫu.
2. Tập mẫu có  $m$  mẫu, mỗi mẫu gồm  $k$  thuộc tính, trong đó có một thuộc tính quyết định. Tổng số  $n$  các giá trị của thuộc tính quyết định chính là số lớp của tập mẫu.
3. Tập luật  $R$  có giá trị khởi tạo là  $\emptyset$
4. Tất cả các dòng trong bảng ban đầu chưa được đánh dấu (kiểm tra).

Áp dụng thuật giải ILA qua các bước:

- *Bước 1:* Chia bảng  $m$  mẫu ban đầu thành  $n$  bảng con. Mỗi bảng con ứng với một giá trị của thuộc tính phân lớp “hay thuộc tính quyết định” của tập mẫu.

(\* thực hiện các bước 2 đến 8 cho mỗi bảng con\*)

- *Bước 2:* Khởi tạo biến đếm kết hợp thuộc tính  $j$ ,  $j = 1$  cho biết số thuộc tính đang xét.
- *Bước 3:* Với mỗi bảng con đang khảo sát, phân chia danh sách các thuộc tính theo các tổ hợp phân biệt, mỗi tổ hợp ứng với  $j$  thuộc tính phân biệt.
- *Bước 4:* Với mỗi tổ hợp các thuộc tính, tính số lượng các giá trị thuộc tính xuất hiện theo cùng tổ hợp thuộc tính trong các dòng chưa được đánh dấu của bảng con đang xét (mà đồng thời không xuất hiện với tổ hợp thuộc tính này trên các bảng còn lại). Gọi tổ hợp đầu tiên (trong bảng con) có số lần xuất hiện nhiều nhất là *tổ hợp lớn nhất*.
- *Bước 5:* Nếu *tổ hợp lớn nhất* bằng  $\emptyset$ , tăng  $j$  lên 1 và quay lại bước 3.
- *Bước 6:* Đánh dấu các dòng thỏa *tổ hợp lớn nhất* của bảng con đang xử lý theo lớp.
- *Bước 7:* Thêm luật mới vào tập luật  $R$ , với vế trái là tập các giá trị của thuộc tính ứng với *tổ hợp lớn nhất* (kết hợp các thuộc tính bằng toán tử AND) và vế phải là giá trị thuộc tính quyết định tương ứng.
- *Bước 8:* Nếu tất cả các dòng đều đã được đánh dấu phân lớp, tiếp tục thực hiện từ bước 2 cho các bảng con còn lại. Ngược lại (nếu chưa đánh dấu hết các dòng) thì quay lại bước 4. Nếu tất cả các bảng con đã được xét thì kết thúc, kết quả thu được là tập luật cần tìm.

Thuật giải ILA được đánh giá mạnh hơn hai thuật giải về phương pháp học quy nạp trước đây là ID3

Ví dụ minh họa thuật giải

Mẫu số	Size	Color	Shape	Decision
1	medium	blue	brick	yes
2	small	red	wedge	no
3	small	red	sphere	yes
4	large	red	wedge	no
5	large	green	pillar	yes
6	large	red	pillar	no
7	large	green	sphere	yes

Bước 1:

*Bảng con 1*

Mẫu số cũ	Mẫu số mới	Size	Color	Shape	Decision
1	1	medium	blue	brick	yes
3	2	small	red	sphere	yes
5	3	large	green	pillar	Yes
7	4	large	green	sphere	Yes

*Bảng con 2*

Mẫu số cũ	Mẫu số mới	Size	Color	Shape	Decision
2	1	small	red	wedge	No
4	2	large	red	wedge	No
6	3	large	red	pillar	No

Thuật giải ILA được đánh giá mạnh hơn hai thuật giải về phương pháp học quy nạp trước đây là ID3

Với  $j=1$ , danh sách các tổ hợp thuộc tính gồm có {Size}, {Color}, và {Shape}.

Với tổ hợp {Size}, giá trị thuộc tính “medium” xuất hiện trong bảng con thứ nhất nhưng không có trong bảng con thứ hai, do đó giá trị tổ hợp lớn nhất là “medium”. Bởi vì các giá trị thuộc tính “small” và “large” xuất hiện trong cả hai bảng con, nên không được xét trong bước này. Với tổ hợp {Size} giá trị thuộc tính “medium” chỉ bằng 1

Xét tiếp cho tổ hợp {Color} thì giá trị tổ hợp lớn nhất là bằng 2, ứng với thuộc tính “green”, còn thuộc tính “blue” là bằng 1

Với tổ hợp {Shape}, ta có “brick” xuất hiện một lần, và “sphere” hai lần.

→ Đến đây nhận thấy có hai thuộc tính: “green” và “sphere” ở hai thuộc tính {Color} và {Shape} đều có số lần xuất hiện lớn nhất là 2. Thuật toán **mặc định chọn trường hợp thứ nhất để xác định luật tổ hợp lớn nhất**. Dòng 3 và 4 được đánh dấu đã phân lớp

<b>R<sub>1</sub></b>	3	large	green	pillar	yes
	4	large	green	sphere	yes

ta có luật dẫn như sau:

**Rule 1: IF color IS green THEN decision IS yes**

các mẫu còn lại (chưa đánh dấu) trong bảng con 1 (tức dòng 1 và 2). Áp dụng tương tự như trên, ta thấy giá trị thuộc tính “medium” của {Size}, “blue” của {Color}, “brick” và “sphere” của {Shape} đều xuất hiện một lần. **Bởi vì số lần xuất hiện này giống nhau, thuật giải áp dụng luật mặc định chọn trường hợp đầu tiên là thuộc tính {Size}**. Ta có thêm luật sau:

**Rule 2: IF size IS medium THEN decision IS yes**

<b>R<sub>2</sub></b>	1	medium	blue	brick	yes
----------------------	---	--------	------	-------	-----

Tiếp tục áp dụng bước 4 đến 8 trên dòng còn lại (tức dòng 2). Giá trị thuộc tính “sphere” của {Shape} xuất hiện một lần, ta có luật thứ ba:

**Rule 3: IF shape IS sphere THEN decision IS yes**

<b>R<sub>3</sub></b>	2	small	red	sphere	yes
----------------------	---	-------	-----	--------	-----

Như vậy, tất cả các dòng trong bảng con 1 đã được đánh dấu, ta chuyển qua xử lý tiếp bảng con 2.

Thuộc tính “wedge” của {Shape} xuất hiện hai lần trong dòng 1 và 2 của bảng con này và không xuất hiện bảng con 1. Đánh dấu các dòng này với luật dẫn thứ tư sau:

**Rule 4: IF shape IS wedge THEN decision IS no**

<b>R<sub>4</sub></b>	1	small	red	wedge	no
	2	large	red	wedge	no

Thuật giải ILA được đánh giá mạnh hơn hai thuật giải về phương pháp học quy nạp trước đây là ID3

ILA tăng  $j = j + 1$  khi bảng con thứ 2 chỉ còn một dòng và tất cả các mẫu của các tổ hợp  $j = 1$  đều xuất hiện trong bảng 1.

Khởi tạo các tổ hợp 2 thuộc tính là {Size và Color}, {Size và Shape}, và {Color và Shape}. Các tổ hợp thứ nhất và thứ ba thoả mãn điều kiện không xuất hiện trong bảng con 1 với các cặp thuộc tính hiện có của dòng này.

mới	Size	Color	Shape	Decision
3	Large	red	pillar	no

Theo luật mặc định, ta chọn luật theo trường hợp thứ nhất. Đánh dấu dòng này, ta có thêm luật dẫn thứ 5:

**Rule 5: IF size IS large AND color IS red THEN decision IS no**

**R<sub>5</sub>** 3 | ~~large~~ | ~~red~~ | pillar | no

Tổng hợp các luật

Rule 1: IF color IS green THEN decision IS yes

Rule 2: IF size IS medium THEN decision IS yes

Rule 3: IF shape IS sphere THEN decision IS yes

Rule 4: IF shape IS wedge THEN decision IS no

Rule 5: IF size IS large AND color IS red THEN decision IS no