Thuật toán ILA

Thuật toán nâng cao (Đại học Thủy lợi)

Thuật toán ILA(Inductive Learning Algorithm) được dung để xác định các luật phân loại cho tập hợp các mẫu học. Thuật giải này thực hiện theo cơ chế lặp, để tìm luật riêng đại diện cho tập mẫu của từng lớp. Sau khi xác định được luật, ILA loại bỏ các mẫu liên quan khỏi tập mẫu, đồng thời them luật mới vào tập luận. Kết qủa có được là một danh sách có thứ tự các luật chứ không là một cây quyết. Các ưu điểm của thuật giải này có thể được trình bày như sau:

-Dạng các luật sẽ phù hợp cho việc khảo sát dữ liệu, mô tả mỗi lớp một cách đơn giản để dễ phân biệt với các lớp khác.

- Tập luật được sắp thứ tự, riêng biết cho phép quan tâm đến một luật tại thời điểm bất kỳ. Khác với việc xử lý luật theo phương pháp cây quyết định, vốn rất phức tạp trong trường hợp các nút cây trở nên khá lớn.

Bước 1: Chia mẫu ban đầu thành n bảng con. Mỗi bảng con ứng với một giá trị của thuộc tính quyết định của tập mẫu

Thực hiện lần lượt các bước từ 2 đến 8 cho mỗi bảng con có được

Bước 2: j=1(j là số thuộc tính của tổ hợp T)

Bước 3: Trên mỗi bảng con đang khảo sát, chia danh sách các thuộc tính thành các tổ hợp khác nhau, mỗi tổ hợp bao gồm j thuộc tính

Bước 4: Với mỗi tổ hợp thuộc tính có được tính số lần giá trị thuộc tính xuất hiện theo cùng tổ hợp thuộc tính trong các dòng còn lại của bảng con đang xét (mà đồng thời không xuất hiện tổ hợp giá trị này trên tất cả các bảng còn lại).

Tổ hợp $T^{\text{Thuộc tính}}_{\text{Giá trị của thuộc tính}}$

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 có giá trị bằng 0, tang j lên 1 và quay lại bước 3.

Bước 6:Loại bỏ các dòng thỏa mãn tổ hợp lớn nhất ra khỏi bảng con đang xử lý

Bước 7: Thêm luật mới vào tập luật R, với vế trái là tập cac thuộc tính của tổ hợp lớn nhất(Két 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 loại bỏ, tiếp tục thực hiện bước 2 cho các bảng còn lại. Ngược lại(nếu còn dòng chưa bị loại bỏ) thì quay lại bước 4. Nếu tất cả các dòng con đã được xét thì kết thúc. Tập R chính là tập luật cần tìm.

2. Minh họa thuật toán:

Minh họa giải thuật ILA cho bảng dữ liệu sau đây:

Tên	Nước tiểu	Giảm cân	Đau lưng	Sốt	Kết quả
An	Vàng	Không	Không	Có	Không bệnh sỏi thận
Cường	Vàng	Không	Không	Không	Không bệnh sỏi thận
Châu	Có vôi	Không	Không	Có	Bệnh sỏi thận
Dung	Có máu	ít	Không	Có	Bệnh sỏi thận
Diễm	Có máu	Nhanh	Có	Có	Bệnh sỏi thận
Hương	Có máu	Nhanh	Có	Không	Không bệnh sỏi thận
Hoa	Có vôi	Nhanh	Có	Không	Bệnh sỏi thận
Phương	Vàng	ít	Không	Có	Không bệnh sỏi thận
My	Vàng	Nhanh	Có	Có	Bệnh sỏi thận
Nhung	Có máu	ít	Có	Có	Bệnh sỏi thận
Thu	Vàng	ít	Có	Không	Bệnh sỏi thận
Thương	Có vôi	ít	Không	Không	Bệnh sỏi thận
Tuấn	Có vôi	Không	Có	Có	Bệnh sỏi thận
Tùng	Có máu	ít	Không	Không	Không bệnh sỏi thận

Chia bảng mẫu thành 2 bảng con bởi 2 loại quyết định: “Bệnh sỏi thận” và “Không bệnh sỏi thận” như sau:

Bảng 1: “Bệnh sỏi thận”

Tên	Nước Tiểu	Giảm Cân	Đau Lưng	Sốt	Kết Quả
Châu	Có vôi	Không	Không	Có	Bệnh Sỏi Thận
Dung	Có máu	Ít	Không	Có	Bệnh Sỏi Thận
Diễm	Có máu	Nhanh	Có	Có	Bệnh Sỏi Thận
Hoa	Có vôi	Nhanh	Có	Không	Bệnh Sỏi Thận
My	Vàng	Nhanh	Có	Có	Bệnh Sỏi Thận
Nhung	Có máu	Ít	Có	Có	Bệnh Sỏi Thận
Thu	Vàng	Ít	Có	Không	Bệnh Sỏi Thận
Thương	Có vôi	Ít	Không	Không	Bệnh Sỏi Thận
Tuấn	Có vôi	Không	Có	Có	Bệnh Sỏi Thận

Bảng 2: “Không bệnh sỏi thận”

Tên	Nước Tiểu	Giảm Cân	Đau Lưng	Sốt	Kết Quả
An	Vàng	Không	Không	Có	Không Bệnh Sỏi Thận
Cường	Vàng	Không	Không	Không	Không Bệnh Sỏi Thận
Hương	Có máu	Nhanh	Có	Không	Không Bệnh Sỏi Thận
Phương	Vàng	Ít	Không	Có	Không Bệnh Sỏi Thận
Tùng	Có máu	Ít	Không	Không	Không Bệnh Sỏi Thận

Bảng 1: "Bệnh sỏi thận"

Tổ hợp: T

Với j=1, có 4 tổ hợp:

- {Nước tiểu}
- {Giảm cân}
- {Đau lưng}
- {Sốt}

- Với tổ hợp {Nước tiểu}: Thuộc tính "Có vôi" xuất hiện 4 lần trong bảng 1 và không xuất hiện trong bảng 2. Thuộc tính "Có máu" và "Vàng" xuất hiện trên cả hai bảng.

  $T^{\text{Nước tiểu}}_{\text{Có vôi}} = 4$

  $T^{\text{Nước tiểu}}_{\text{Có máu}} = 0$

  $T^{\text{Nước tiểu}}_{\text{Vàng}} = 0$

- Với tổ hợp {Giảm cân}: Thuộc tính "Không", "Nhanh", "Ít" xuất hiện trên cả hai bảng.

  $T^{\text{Giảm cân}}_{\text{Không}} = 0$

  $T^{\text{Giảm cân}}_{\text{Nhanh}} = 0$

  $T^{\text{Giảm cân}}_{\text{Ít}} = 0$

- Với tổ hợp {Đau lưng}: Thuộc tính "Không", "Có" xuất hiện trên cả hai bảng.

  $T^{\text{Đau lưng}}_{\text{Không}} = 0$

  $T^{\text{Đau lưng}}_{\text{Có}} = 0$

- Với tổ hợp {Sốt}: Thuộc tính "Không", "Có" xuất hiện trên cả hai bảng.

$T^{Sốt}_{Không} = 0$

$T^{Sốt}_{Có} = 0$

⇨ Ta có $T^{Nước\ tiểu}_{Có\ vôi} = 4$ là lớn nhất. Ta chọn $T^{Nước\ tiểu}_{Có\ vôi}$

RULE 1: IF Nước tiểu = có vôi THEN Kết quả = Bệnh sỏi thận

Tiếp theo ta loại bỏ những dòng thỏa mãn tổ hợp lớn nhất tương ứng với Nước tiểu = Có vôi ra khỏi bảng 1 ta có bảng sau:

Tên	Nước tiểu	Giảm cân	Đau lưng	Sốt	Kết quả
Dung	Có máu	Ít	Không	Có	Bệnh sỏi thận
Diễm	Có máu	Nhanh	Có	Có	Bệnh sỏi thận
My	Vàng	Nhanh	Có	Có	Bệnh sỏi thận
Nhung	Có máu	Ít	Có	Có	Bệnh sỏi thận
Thu	Vàng	Ít	Có	Không	Bệnh sỏi thận

Các dòng trong bảng trên mọi giá trị của thuộc tính đều xuất hiện trong cả hai bảng(mọi giá trị T đều bằng 0) nên ta sẽ tang j lên 1.

Với j=2, có 6 tổ hợp:

- {Nước tiểu, Giảm cân }
- {Nước tiểu, Đau lưng}
- {Nước tiểu, Sốt}
- {Giảm cân, Đau lưng}
- {Giảm cân, Sốt}
- {Đau lưng, Sốt}

$T^{Nước\ tiểu\ Giảm\ cân}_{Có\ máu\ ,\ Ít} = 0$

$T^{Nước\ tiểu\ Giảm\ cân}_{Có\ máu\ ,\ Nhanh} = 0$

$T^{Nước\ tiểu\ Giảm\ cân}_{Vàng\ ,\ Ít} = 0$

$T^{Nước\ tiểu\ Giảm\ cân}_{Vàng\ ,\ Nhanh} =$

$T^{Nước\ tiểu\ Đau\ lưng}_{Có\ máu\ ,\ Không} = 0$

$T^{Nước\ tiểu\ Đau\ lưng}_{Có\ máu\ ,\ Có} = 0$

$T^{\text{Nước tiểu}}_{\text{Vàng}}{}^{,}{}^{\text{Đau lưng}}_{\text{Có}} = 1$

$T^{\text{Nước tiểu}}_{\text{Vàng}}{}^{,}{}^{\text{Sốt}}_{\text{Không}} = 0$

$T^{\text{Nước tiểu}}_{\text{Có máu}}{}^{,}{}^{\text{Sốt}}_{\text{Có}} = 3$

$T^{\text{Giảm cân}}_{\text{Ít}}{}^{,}{}^{\text{Đau lưng}}_{\text{Không}} = 0$

$T^{\text{Giảm cân}}_{\text{Nhanh}}{}^{,}{}^{\text{Đau lưng}}_{\text{Có}} = 0$

$T^{\text{Giảm cân}}_{\text{Ít}}{}^{,}{}^{\text{Đau lưng}}_{\text{Có}} = 2$

$T^{\text{Giảm cân}}_{\text{Ít}}{}^{,}{}^{\text{Sốt}}_{\text{Có}} = 0$

$T^{\text{Giảm cân}}_{\text{Ít}}{}^{,}{}^{\text{Sốt}}_{\text{Không}} = 0$

$T^{\text{Giảm cân}}_{\text{Nhanh}}{}^{,}{}^{\text{Sốt}}_{\text{Có}} = 1$

$T^{\text{Đau lưng}}_{\text{Không}}{}^{,}{}^{\text{Sốt}}_{\text{Có}} = 0$

$T^{\text{Đau lưng}}_{\text{Có}}{}^{,}{}^{\text{Sốt}}_{\text{Có}} = 2$

$T^{\text{Đau lưng}}_{\text{Có}}{}^{,}{}^{\text{Sốt}}_{\text{Không}} = 0$

⇨ Ta có $T^{\text{Nước tiểu}}_{\text{Có máu}}{}^{,}{}^{\text{Sốt}}_{\text{Có}} = 3$ là lớn nhất, ta chọn $T^{\text{Nước tiểu}}_{\text{Có máu}}{}^{,}{}^{\text{Sốt}}_{\text{Có}}$

RULE 2: IF Nước tiểu= Có máu AND Sốt = Có THEN Kết Quả = Bệnh sỏi thận

Kể tiếp, loại bỏ những dòng ứng với Nước tiểu = có máu và sốt = có ra khỏi bảng ta được:

Tên	Nước tiểu	Giảm cân	Đau lưng	Sốt	Kết quả
My	Vàng	Nhanh	Có	Có	Bệnh sỏi thận
Thu	Vàng	Ít	Có	Không	Bệnh sỏi thận

Với j=2, ta có 6 tổ hợp mỗi tổ hợp gồm 2 thuộc tính:

- {Nước tiểu, Giảm cân }
- {Nước tiểu, Đau lưng}
- {Nước tiểu, Sốt}
- {Giảm cân, Đau lưng}
- {Giảm cân, Sốt}
- {Đau lưng, Sốt}
-

$T^{\text{Nước tiểu}}_{\text{vàng}}, {}^{\text{giảm cân}}_{\text{ít}}=0$

$T^{\text{Nước tiểu}}_{\text{vàng}}, {}^{\text{giảm cân}}_{\text{Nhanh}}=1$

$T^{\text{Nước tiểu}}_{\text{vàng}}, {}^{\text{đau lưng}}_{\text{có}}=2$

$T^{\text{Nước tiểu}}_{\text{vàng}}, {}^{\text{sốt}}_{\text{không}}=0$

$T^{\text{Nước tiểu}}_{\text{vàng}}, {}^{\text{sốt}}_{\text{Có}}=0$

$T^{\text{Giảm cân}}_{\text{ít}}, {}^{\text{đau lưng}}_{\text{có}}=1$

$T^{\text{Giảm cân}}_{\text{ít}}, {}^{\text{sốt}}_{\text{Không}}=0$

$T^{\text{Giảm cân}}_{\text{Nhanh}}, {}^{\text{sốt}}_{\text{Có}}=1$

$T^{\text{Đau lưng}}_{\text{có}}, {}^{\text{sốt}}_{\text{không}}=0$

$T^{\text{Đau lưng}}_{\text{có}}, {}^{\text{sốt}}_{\text{có}}=1$

⇨ Ta có $T^{\text{Nước tiểu}}_{\text{vàng}}, {}^{\text{đau lưng}}_{\text{có}}=2$ là lớn nhất.Ta chọn $T^{\text{Nước tiểu}}_{\text{vàng}}, {}^{\text{đau lưng}}_{\text{có}}$ và ta có luật:

RULE 3: IF Nước tiểu = Vàng AND Đau lưng=Có THEN Kết quả=Bệnh sỏi thận

Loại bỏ các dòng tương ứng với Nước tiểu = vàng, đau lưng=có, như vậy tất cả các dòng trong bảng 1 bị loại bỏ

Bảng 2: "Không bệnh sỏi thận"

Tên	Nước Tiểu	Giảm Cân	Đau Lưng	Sốt	Kết Quả
An	Vàng	Không	Không	Có	Không Bệnh Sỏi Thận
Cường	Vàng	Không	Không	Không	Không Bệnh Sỏi Thận
Hương	Có máu	Nhanh	Có	Không	Không Bệnh Sỏi Thận
Phương	Vàng	Ít	Không	Có	Không Bệnh Sỏi Thận
Tùng	Có máu	Ít	Không	Không	Không Bệnh Sỏi Thận

Trong bảng 2, mọi giá trị của thuộc tính đều xuất hiện trong cả hai bảng(mọi giá trị T đều bằng 0) nên ta sẽ tang j lên 1.

Với j=2, có 6 tổ hợp mỗi tổ hợp có 2 thuộc tính:

- {Nước tiểu, Giảm cân }
- {Nước tiểu, Đau lưng}
- {Nước tiểu, Sốt}
- {Giảm cân, Đau lưng}
- {Giảm cân, Sốt}
- {Đau lưng, Sốt}

$T(^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{giảm cân}}_{\text{ít}})=0$

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{giảm cân}}_{\text{nhanh}} =0$

$T^{\text{nước tiểu}}_{\text{vàng}}, ^{\text{giảm cân}}_{\text{ít}} = 0$

$T^{\text{nước tiểu}}_{\text{vàng}}, ^{\text{giảm cân}}_{\text{không}}=2$

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{đau lưng}}_{\text{không}}=0$

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{đau lưng}}_{\text{có}} =0$

$T^{\text{nước tiểu}}_{\text{vàng}}, ^{\text{đau lưng}}_{\text{không}}=3$

$T^{\text{nước tiểu}}_{\text{vàng}}, ^{\text{sốt}}_{\text{không}} =0$

$T^{\text{nước tiểu}}_{\text{vàng}}, ^{\text{sốt}}_{\text{có}} =0$

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{sốt}}_{\text{không}}=2$

$T^{\text{giảm cân}}_{\text{ít}}, ^{\text{đau lưng}}_{\text{không}}=0$

$T^{\text{Giảm cân}}_{\text{nhanh}}, ^{\text{đau lưng}}_{\text{có}}=0$

$T^{\text{giảm cân}}_{\text{không}}, ^{\text{đau lưng}}_{\text{không}}=0$

$T^{\text{giảm cân}}_{\text{không}}, ^{\text{sốt}}_{\text{có}}=0$

$T^{\text{giảm cân}}_{\text{không}}, ^{\text{sốt}}_{\text{không}} =1$

$T^{\text{giảm cân}}_{\text{nhanh}}, ^{\text{sốt}}_{\text{không}}=0$

$T^{\text{giảm cân}}_{\text{ít}}, ^{\text{sốt}}_{\text{có}}=0$

$T^{\text{giảm cân}}_{\text{ít}}, ^{\text{sốt}}_{\text{không}} = 0$

$T^{\text{đau lưng}}_{\text{không}}, ^{\text{sốt}}_{\text{có}} = 0$

$T^{\text{đau lưng}}_{\text{không}}, ^{\text{sốt}}_{\text{không}} = 0$

$T^{\text{đau lưng}}_{\text{có}}, ^{\text{sốt}}_{\text{không}} = 0$

⇨ Ta có $T^{\text{nước tiểu}}_{\text{vàng}}, ^{\text{đau lưng}}_{\text{không}} = 3$ là lớn nhất. ta chọn $T^{\text{nước tiểu}}_{\text{vàng}}, ^{\text{đau lưng}}_{\text{không}}$ và ta có luật

RULE 4: IF Nước tiểu = Vàng AND Đau lưng = Không THEN Kết Quả = Không bệnh sỏi thận

Kết tiếp ta loại bỏ những dòng ứng với Nước tiểu=vàng và Đau lưng = Không ra khỏi Bảng 2:

Tên	Nước tiểu	Giảm cân	Đau lưng	Sốt	Kết quả
Hương	Có máu	Nhanh	Có	Không	Không bệnh sỏi thận
Tùng	Có máu	Ít	Không	Không	Không bệnh sỏi thận

Với j=2, có 6 tổ hợp gồm 2 thuộc tính:

- {Nước tiểu, Giảm cân }
- {Nước tiểu, Đau lưng}
- {Nước tiểu, Sốt}
- {Giảm cân, Đau lưng}
- {Giảm cân, Sốt}
- {Đau lưng, Sốt}

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{giảm cân}}_{\text{ít}} = 0$

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{giảm cân}}_{\text{nhanh}} = 0$

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{đau lưng}}_{\text{không}} = 0$

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{đau lưng}}_{\text{có}} = 0$

$T^{\text{nước tiểu}}_{\text{có máu}}, ^{\text{sốt}}_{\text{không}} = 2$

$T^{\text{giảm cân}}_{\text{ít}}, ^{\text{đau lưng}}_{\text{không}} = 0$

$T^{\text{giảm cân}}_{\text{nhanh}}, ^{\text{đau lưng}}_{\text{có}} = 0$

$T^{\text{giảm cân}}_{\text{nhanh}}, ^{\text{sốt}}_{\text{không}} = 0$

$T^{\text{giảm cân}}_{\text{ít}}, {}^{\text{sốt}}_{\text{không}} = 0$

$T^{\text{đau lưng}}_{\text{không}}, {}^{\text{sốt}}_{\text{không}} = 0$

$T^{\text{đau lưng}}_{\text{có}}, {}^{\text{sốt}}_{\text{không}} = 0$

⇨ Ta có $T^{\text{nước tiểu}}_{\text{có máu}}, {}^{\text{sốt}}_{\text{không}} = 2$ là lớn nhất. ta chọn $T^{\text{nước tiểu}}_{\text{có máu}}, {}^{\text{sốt}}_{\text{không}}$ và ta sẽ có luật:

RULE 5: IF Nước tiểu=Có máu AND Sốt = Không THEN Kết Quả=Không bệnh sỏi thận

Kết tiếp, loại bỏ những dòng ứng với Nước tiểu= có máu và sốt = không ra khỏi bảng

⇨ Như vậy ta đã loại bỏ tất cả các dòng trong bảng 2

⇨ **Thuật toán kết thúc vì tất cả các bảng đã được xét đến và các dòng trong các bảng đã được loại bỏ**.

Tổng hợp các luật:

RULE 1: IF Nước tiểu = có vôi THEN Kết quả = Bệnh sỏi thận

RULE 2: IF Nước tiểu= Có máu AND Sốt = Có THEN Kết Quả = Bệnh sỏi thận

RULE 3: IF Nước tiểu = Vàng AND Đau lưng=Có THEN Kết quả=Bệnh sỏi thận

RULE 4: IF Nước tiểu = Vàng AND Đau lưng = Không THEN Kết Quả = Không bệnh sỏi thận

RULE 5: IF Nước tiểu=Có máu AND Sốt = Không THEN Kết Quả=Không bệnh sỏi thận

3. Cài đặt ứng dụng minh họa

Hai chuyên đề nổi bật là giải thuật để xây dựng cây định danh và tìm ra tri thức cho

mẫu dữ liệu thực tế là giải thuật

Quinlan và giải thuật ILA. Trong phần này ứng dụng

chỉ minh họa cho giải thuật ILA để tìm ra tri thức cho bảng dữ liệu.