### 

### 

BÁO CÁO NỘI DUNG PHÂN LỚP VỚI CÂY QUYẾT ĐỊNH - Nhóm 1

### 

### 

### 

### 

### 

TP HỒ CHÍ MINH, THÁNG 8 NĂM 2018

MÔN Khai thác dữ liệu

Bộ môn Hệ thống thông tin – Khoa Công nghệ thông tin – ĐH Công nghiệp TP. HCM

### 

Mục lục

[I. Tổng quan 3](#_slmr1hnwo9bh)

[1. Giới thiệu về phân lớp (Classification) 3](#_xzaqupw3xqyc)

[2. Phân loại một số cách phân loại dữ liệu 3](#_v1do3kb8dqm2)

[II. Ứng dụng của việc phân lớp dữ liệu 5](#_nj4om0cef4gq)

[1. Ứng dụng trong lĩnh vực tài chính 5](#_yx854rv6fmxc)

[2. Ứng dụng trong lĩnh vực giáo dục 5](#_av8nwddllb57)

[III. Phương pháp cây quyết định (Decision Tree) 6](#_kc52ate08m14)

[1. Cấu trúc cây quyết định 6](#_mjxki7fyo8wu)

[2. Xây dựng cây 6](#_4ctvb5xjgran)

[3. Cắt tỉa cây (Pruning) 8](#_2yxym5wug94f)

[4. Đánh giá một số cách phổ biến để xây dựng cây quyết định: 8](#_d3m3leecwhy0)

[IV. Phương pháp dựa trên luật (Rule-Based Classification) 10](#_k2bkd3qnwwx4)

[1. Cấu trúc của luật 10](#_8werdahglq6o)

[2. Quá trình xây dựng luật 10](#_1ezdpx28vngl)

[3. Ưu nhược điểm 10](#_xk8i41l4njbo)

# 

### Tổng quan

# Giới thiệu về phân lớp (Classification)

* 1. Phân lớp là gì ?
* Classification là một nhiệm vụ trong khai thác dữ liệu liên quan đến việc gán nhãn lớp cho từng trường hợp trong một tập dữ liệu dựa trên các tính năng của nó.
  1. Mục tiêu của nó ?
* Mục tiêu là xây dựng một mô hình dự đoán chính xác nhãn lớp của các trường hợp mới dựa trên các tính năng của chúng.

Ví dụ về Nhiệm vụ Classification:

* Dự đoán tế bào khối u là lành tính hay ác tính
* Phân loại giao dịch thẻ tín dụng là hợp pháp hay gian lận
* Phân loại tin tức thành tài chính, thời tiết, giải trí, thể thao, v.v.
* Xác định email rác, trang web rác, nội dung dành cho người lớn
* Hiểu được truy vấn web có mục đích thương mại hay không

# Phân loại một số cách phân loại dữ liệu

* 1. Decision Tree Induction
* Decision tree sử dụng cấu trúc cây để phân loại dữ liệu dựa trên các thuộc tính (features) của nó. Mỗi nút trong cây đại diện cho một thuộc tính, và mỗi nhánh đại diện cho một giá trị của thuộc tính đó.

Ưu điểm:

* Dễ hiểu và trực quan.
* Dễ áp dụng cho cả dữ liệu số và dữ liệu danh mục.
* Có thể xử lý dữ liệu thiếu và không yêu cầu chuẩn hóa dữ liệu.

Nhược điểm:

* Dễ dẫn đến quá khớp (overfitting).
* Hiệu quả có thể không cao nếu dữ liệu phức tạp.
  1. Bayes Classification Methods
* Dựa trên định lý Bayes, các thuật toán phân loại Bayes sử dụng xác suất có điều kiện để dự đoán lớp của đối tượng.
* Naive Bayes: Giả định rằng các đặc trưng là độc lập với nhau. Đây là một phương pháp phân loại đơn giản và hiệu quả.

Ưu điểm:

* Nhanh chóng, yêu cầu ít tài nguyên tính toán.
* Hiệu quả với dữ liệu lớn và trường hợp nhiều lớp.

Nhược điểm:

* Giả định về tính độc lập của các thuộc tính có thể không đúng trong thực tế.
* Hiệu quả giảm nếu dữ liệu không tuân theo phân phối xác suất giả định.
  1. Rule-Based Classification
* Các thuật toán phân loại dựa trên luật tạo ra các luật dưới dạng "If-Then" để phân loại dữ liệu. Mỗi luật chỉ ra điều kiện khi một đối tượng thuộc về một lớp cụ thể.

Ưu điểm:

* Kết quả dễ hiểu và giải thích.
* Phù hợp khi cần sự rõ ràng và dễ dàng áp dụng vào hệ thống chuyên gia.

Nhược điểm:

* Quá trình tìm kiếm luật hiệu quả có thể phức tạp.
* Dễ gặp tình trạng quá khớp nếu không điều chỉnh tốt.
  1. Other techniques

Support Vector Machines (SVM): Phân loại dữ liệu bằng cách tìm một siêu phẳng tối ưu để phân tách các điểm dữ liệu thuộc các lớp khác nhau.

K-Nearest Neighbors (KNN): Phân loại dựa trên khoảng cách giữa điểm cần phân loại và các điểm dữ liệu lân cận nó trong tập huấn luyện.

### 

### Ứng dụng của việc phân lớp dữ liệu

### 1. Ứng dụng trong lĩnh vực tài chính

* Phân tích tín dụng: Phân lớp được sử dụng để dự đoán khả năng trả nợ của khách hàng dựa trên các thông tin cá nhân và tài chính của họ. Các ngân hàng sử dụng phân loại để phân nhóm khách hàng thành những người có nguy cơ nợ xấu và những người có khả năng thanh toán tốt.
  + Ví dụ: Sử dụng phương pháp phân lớp để phân chia khách hàng thành "Rủi ro cao", "Rủi ro thấp" và "Không có rủi ro" khi xét duyệt hồ sơ vay vốn.
* Phát hiện gian lận: Phân lớp giúp phát hiện các giao dịch gian lận trong thẻ tín dụng bằng cách phân loại các giao dịch thành "Hợp pháp" và "Gian lận" dựa trên hành vi giao dịch.
  + Ví dụ: Sử dụng các thuật toán Bayes hoặc mạng nơ-ron để phân loại các giao dịch tín dụng trong thời gian thực.

### 2. Ứng dụng trong lĩnh vực giáo dục

* Phân loại kết quả học tập: Phân loại sinh viên thành các nhóm dựa trên kết quả học tập và hoạt động học tập để hỗ trợ giảng dạy cá nhân hóa.
  + Ví dụ: Sử dụng các thuật toán phân loại để xác định học sinh có nguy cơ trượt môn, từ đó đưa ra các biện pháp can thiệp.
* Phân loại tài liệu học tập: Hệ thống phân lớp giúp tổ chức và phân loại tài liệu học tập như bài giảng, sách, và bài kiểm tra.
  + Ví dụ: Phân loại tài liệu học tập thành các chủ đề hoặc độ khó khác nhau trong các nền tảng học trực tuyến.

### Phương pháp cây quyết định (Decision Tree)

# Cấu trúc cây quyết định

1. Cây quyết định là một mô hình hình cây sử dụng để phân loại và hồi quy. Cấu trúc của cây quyết định bao gồm:
   * Nút gốc (Root Node): Nút đầu tiên của cây, đại diện cho toàn bộ tập dữ liệu.
   * Nút nội (Internal Nodes): Các nút trong cây dùng để phân chia tập dữ liệu thành các nhánh dựa trên thuộc tính.
   * Nút lá (Leaf Nodes): Các nút cuối cùng không có bất kỳ phân chia nào. Mỗi nút lá đại diện cho một lớp hoặc giá trị dự đoán.

# Xây dựng cây

* Chọn thuộc tính: Sử dụng các tiêu chí như Entropy hoặc Gini Index để chọn thuộc tính tốt nhất để phân chia dữ liệu.
* Tạo nút: Tạo nút mới cho thuộc tính đã chọn và phân chia dữ liệu thành các nhánh dựa trên các giá trị của thuộc tính đó.
* Lặp lại: Lặp lại quá trình cho các nhánh cho đến khi đạt được một số điều kiện dừng, chẳng hạn như không còn thuộc tính nào để phân chia hoặc tất cả các nhánh đều thuộc về cùng một lớp.

Một số cách phổ biến để xây dựng cây quyết định:

1. Information Gain

Công Thức Entropy ()

Trong đó:

* là tập dữ liệu.
* là số lớp (labels) trong dữ liệu.
* ​ là xác suất xuất hiện của lớp thứ iii trong tập dữ liệu .

Công Thức Entropy Của Thuộc Tính ()

Trong đó:

* là số lượng giá trị khác nhau của thuộc tính .
* là tập con của , bao gồm các đối tượng có giá trị của thuộc tính là .
* ​ là tỷ lệ của tập con so với tập dữ liệu gốc .
* là entropy của tập con .

Công Thức Information Gain ()

Trong đó:

* là entropy ban đầu của tập dữ liệu
* là entropy của tập dữ liệu sau khi được phân chia theo thuộc tính .

1. Gain Ratio(C4.5)

Công Thức Tính SplitInfo:

Trong đó:

* là tập dữ liệu..
* là số lượng giá trị khác nhau của thuộc tính .
* là tập con của , bao gồm các đối tượng có giá trị của thuộc tính là .

Công thức tính Gain Ratio

Trong đó:

* là Information Gain của thuộc tính .
* là độ phân tán dữ liệu sau khi chia.

1. Gini Index

Công Thức Tính Chỉ Số Gini:

* Nếu tập dữ liệu được chia theo thuộc tính thành hai tập con và , chỉ số Gini của việc chia dữ liệu là:

Trong đó:

* là tần suất tương đối của lớp trong tập dữ liệu .
* Gini Index đo lường mức độ không thuần nhất của dữ liệu, càng nhỏ thì dữ liệu càng thuần nhất.

Chỉ Số Gini Sau Khi Chia Tập Dữ Liệu:

=

Giảm Độ Không Thuần Nhất:

* Sự giảm độ không thuần nhất sau khi chia dữ liệu được tính bằng:

# Cắt tỉa cây (Pruning)

* Cắt tỉa là quá trình loại bỏ các nhánh không cần thiết trong cây quyết định để giảm thiểu hiện tượng overfitting. Có hai loại cắt tỉa:
* Cắt tỉa tiền xử lý (Pre-pruning): Ngừng xây dựng cây khi độ sâu hoặc số lượng mẫu trong nút giảm xuống dưới một ngưỡng nhất định.
* Cắt tỉa hậu xử lý (Post-pruning): Xây dựng cây đầy đủ trước và sau đó cắt tỉa các nút lá dựa trên hiệu suất của cây trên tập dữ liệu kiểm tra.

# Đánh giá một số cách phổ biến để xây dựng cây quyết định:

1. Information Gain (ID3):

* Ưu điểm: Cho kết quả tốt trong nhiều trường hợp.
* Nhược điểm: Thiên vị các thuộc tính có nhiều giá trị (multivalued attributes), có thể dẫn đến việc lựa chọn các thuộc tính không thực sự có giá trị phân loại cao.

2. Gain Ratio (C4.5):

* Ưu điểm: Được sử dụng để khắc phục sự thiên vị của Information Gain, ưa chuộng các thuộc tính có sự phân chia không cân bằng.
* Nhược điểm: Có xu hướng ưu tiên các phép chia không cân bằng, khi mà một phân vùng nhỏ hơn đáng kể so với các phân vùng khác.

3. Gini Index (CART):

* Ưu điểm: Tập trung vào việc giữ độ thuần nhất của các tập con sau khi phân chia.
* Nhược điểm:
  + Thiên vị các thuộc tính có nhiều giá trị.
  + Gặp khó khăn khi số lượng lớp lớn.
  + Có xu hướng ưa chuộng các phép kiểm tra tạo ra các phân vùng có kích thước bằng nhau và độ thuần khiết trong cả hai phân vùng.

### Phương pháp dựa trên luật (Rule-Based Classification)

# Cấu trúc của luật

* Quy tắc điều kiện-hành động: Các quy tắc thường được định dạng dưới dạng "Nếu ... thì ...".

Ví dụ:

* IF (Blood Type=Warm) (Lay Eggs=Yes) Birds
* (Taxable Income < 50K) (Refund=Yes) Evade=No
* IF (age = youth) AND (student = yes) THEN (buys\_computer = yes)

# Quá trình xây dựng luật

* Thu thập thông tin: Tập hợp dữ liệu và xác định các thuộc tính cần thiết.
* Xác định quy tắc: Dựa trên kiến thức chuyên môn và phân tích dữ liệu, xác định các quy tắc điều kiện.
* Kiểm tra và điều chỉnh: Áp dụng quy tắc trên dữ liệu kiểm tra để đánh giá hiệu suất và điều chỉnh quy tắc khi cần thiết.

1. Cách thức hoạt động

* Một luật phủ một mẫu nếu các thuộc tính của mẫu thỏa mãn điều kiện của luật.

Tập hợp các luật có thể là:

* Loại trừ lẫn nhau: Mỗi mẫu chỉ được phủ bởi tối đa một luật.
* Bao quát toàn bộ: Mỗi mẫu được phủ bởi ít nhất một luật.

1. Các chỉ số đánh giá
   1. Độ bao phủ (Coverage): Tỷ lệ các mẫu thỏa mãn điều kiện của luật

Phần các bản ghi đáp ứng tiền đề của một quy tắc

ncovers: Là số bộ dữ liệu được bao phủ bởi quy tắc R.

ncorrect: Là số bộ dữ liệu được phân loại chính xác bởi quy tắc .

* 1. Độ chính xác (Accuracy): Tỷ lệ các mẫu vừa thỏa điều kiện vừa đúng với nhãn lớp mà luật dự đoán.

1. Các kỹ thuật sinh luật
   1. Phương pháp trực tiếp: Trích xuất luật trực tiếp từ tập dữ liệu (VD: các thuật toán ILA, CN2, RIPPER).
   2. Phương pháp gián tiếp: Trích xuất luật từ các mô hình khác như cây quyết định hoặc mạng nơron.
2. Thuật toán bao phủ tuần tự (Sequential Covering Algorithm)

* Là phương pháp phổ biến để sinh tập hợp luật.

Các bước:

B1 Bắt đầu với một luật rỗng.

B2. Phát triển luật bằng cách thêm điều kiện.

B3.Loại bỏ các mẫu được luật bao phủ.

B4. Lặp lại cho đến khi tất cả các mẫu được bao phủ hoặc đạt điều kiện dừng.

1. Xếp thứ tự luật

* Dựa trên kích thước: Ưu tiên các luật có nhiều kiểm tra thuộc tính hơn.
* Dựa trên lớp: Xếp thứ tự luật theo tỷ lệ xuất hiện của các lớp hoặc chi phí sai lệch phân loại.
* Dựa trên luật: Luật được tổ chức theo chất lượng hoặc do các chuyên gia định nghĩa.

1. Giải quyết xung đột

* Nếu nhiều luật được kích hoạt, cần sử dụng các phương pháp giải quyết xung đột như xếp thứ tự theo kích thước hoặc xếp thứ tự theo lớp để chọn ra luật tốt nhất.

# Ưu nhược điểm

Ưu điểm:

* Dễ giải thích: Các quy tắc rõ ràng và dễ hiểu, giúp người dùng dễ dàng theo dõi quyết định của hệ thống.
* Thích hợp cho các trường hợp rõ ràng: Hiệu quả trong các tình huống mà quy tắc có thể xác định một cách dễ dàng.

Nhược điểm:

* Không linh hoạt: Các quy tắc không thể tự động cập nhật khi dữ liệu mới xuất hiện, cần phải có sự can thiệp của con người.
* Khó quản lý: Khi số lượng quy tắc trở nên lớn, việc quản lý và điều chỉnh chúng có thể trở nên phức tạp.

1. Ví dụ về phân lớp dữ liệu Decision Tree



Phương pháp Gain :

|S| =8 , |P mưa| = 4 , |P không mưa| = 4

Value( Mây) = Ít , Nhiều

- Info(S) = I(4,4) = -4/8\*log2\* 4/8 - ( 4/8\*log2\*4/8) = 1

- Info(S Mây, Ít) = I(0,3) = 0

- Info(S Mây,Nhiều) = I(4,1) = -4/5 \*log2\*4/5 - (1/5\*log2\*1/5) = 0.722

- Info(S Mây) = 3/8 \* 0 + 5/8 \* 0.772 = 0.451

- IG( S Mây) = Info(S) - Info(S Mây) = 1- 0.451 = 0.549

Value( Áp suất ) = Cao , Thấp , Trung Bình

- Info( S áp suất, cao) = I(2,2) = -2\*log2\*2 -( 2\*log2\*2) = 1

- Info( S áp suất , thấp ) = I( 2,1) = -2\*log2 \*2 -( 1 \*log2 \*1) =0.918

- Info( S áp suất , trung bình) = I( 1,0) = 0

- Info(S áp suất ) = 4/8 \* 1 + 3/8\*0.918 +1/8\*0 = 0.844

- IG( S áp suất) = 1- 0.844 = 0.155

Value(Gió) = Bắc , Nam

- Info(Gió , Bắc) = I(3,2) = -3 \*log2\* 2 -( 2\* log2 \*2) =0.971

- Info(Gió, Nam) = I(1,2) = -1\*log2\*1- ( 2\*log2\*2)= 0.918

- Info(Gió) = 5/8\*0.917 + 3/8 \* 0.918 = 0.95

- IG(Gió) = 1 – 0.95 = 0.05

Dựa vào giá trị trên ta chọn Max là IG(Mây) làm node gốc .

Với thuộc tính "Mây" được chọn làm nút đầu tiên, chúng ta sẽ phân chia dữ liệu dựa trên giá trị "Ít" và "Nhiều".

Value (Mây = “ít) :

- Info(Mây ,ít) = I(0,3) = 0 ( Không mưa)

Value ( Mây = “nhiều”)

- Info(Mây, nhiều) = I(4,1) = 0.722 ( Mưa)

Tiếp tục phân chia nhánh “Mây” = Nhiều

- IG(S, áp suất ) = 0.155

- IG(S,Gió) = 0.05

- IG(S,áp suất) > IG(S,Gió) = > chọn IG(S,áp suất) để tiếp tục .

Phân chia nhánh "Nhiều" dựa trên giá trị của "Áp suất"

- If “ áp suất” = “ Cao” = I(2,0) = 0 ( Mưa)

- If “áp suất” = “ Thấp” = I(1,1) = -1\*log2\*1 - ( 1\*log2\*1) = 1

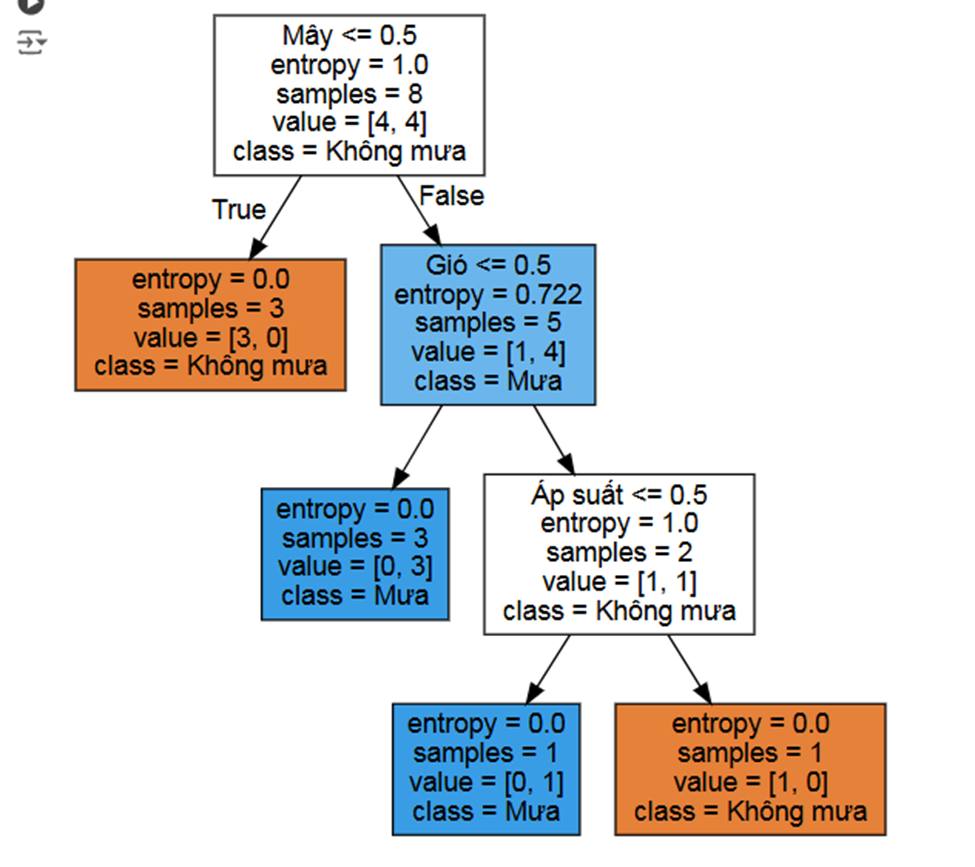
- If “áp suất” = “Trung bình” = I(1,0) = 0 (Mưa)

Tiếp tục phân nhánh với áp suất = “ thấp” dựa trên gió :

- If “ Gió “ = “Bắc” = I(1,0) = 0 ( Mưa)

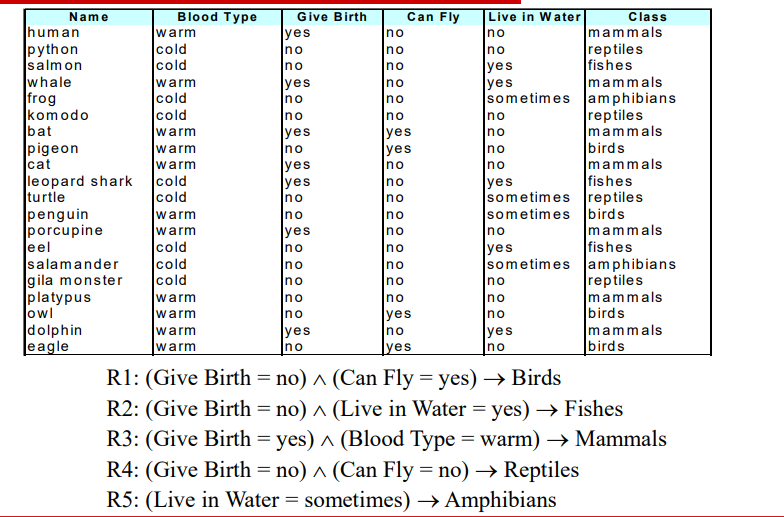
- If “Gió” = “Nam” = I(0,1) = 0 ( Không mưa)

Sơ đồ cây :



1. Ví dụ về phân lớp dữ liệu Rule-Based

* Ví dụ :Phân loại dựa trên quy tắc



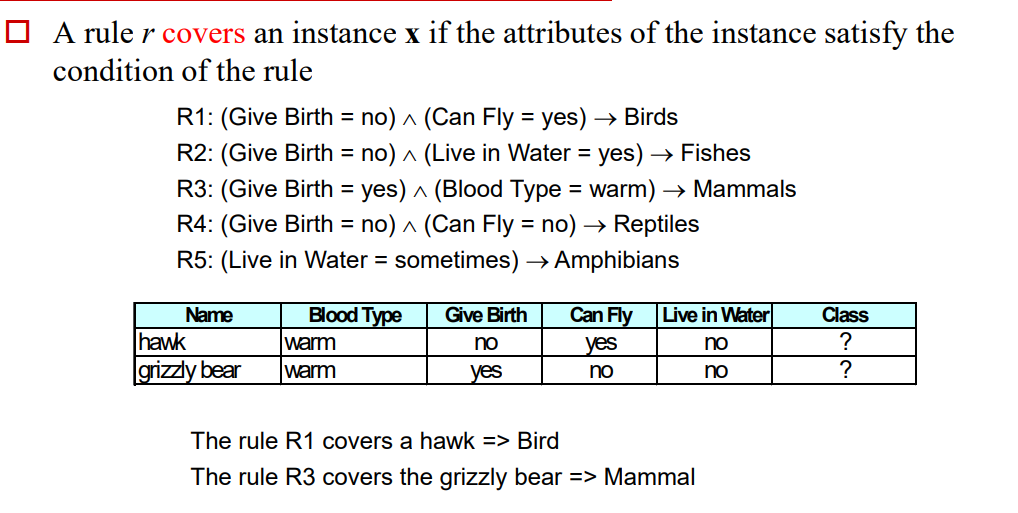
- R1: (Đẻ con = không) ^ (Biết bay = có) -> Chim

- R2: (Đẻ con = không) ^ (Sống dưới nước = có) -> Cá

- R3: (Đẻ con = có) ^ (Nhóm máu = ấm) -> Động vật có vú

- R4: (Đẻ con = không) ^ (Biết bay = không) -> Bò sát

- R5: (Sống dưới nước = đôi khi) -> Lưỡng cư



- Quy tắc r bao phủ một thể hiện x nếu các thuộc tính của thể hiện thỏa mãn điều kiện của quy tắc

R1: (Give Birth = no) ^ (Can Fly = yes) => Birds

R2: (Give Birth = no) ^ (Live in Water = yes) => Fishes

R3: (Give Birth = yes) ^ (Blood Type = warm) => Mammals

R4: (Give Birth = no) ^ (Can Fly = no) => Reptiles

R5: (Live in Water = sometimes) => Amphibians

Phân loại theo quy tắc hoạt động như thế nào?

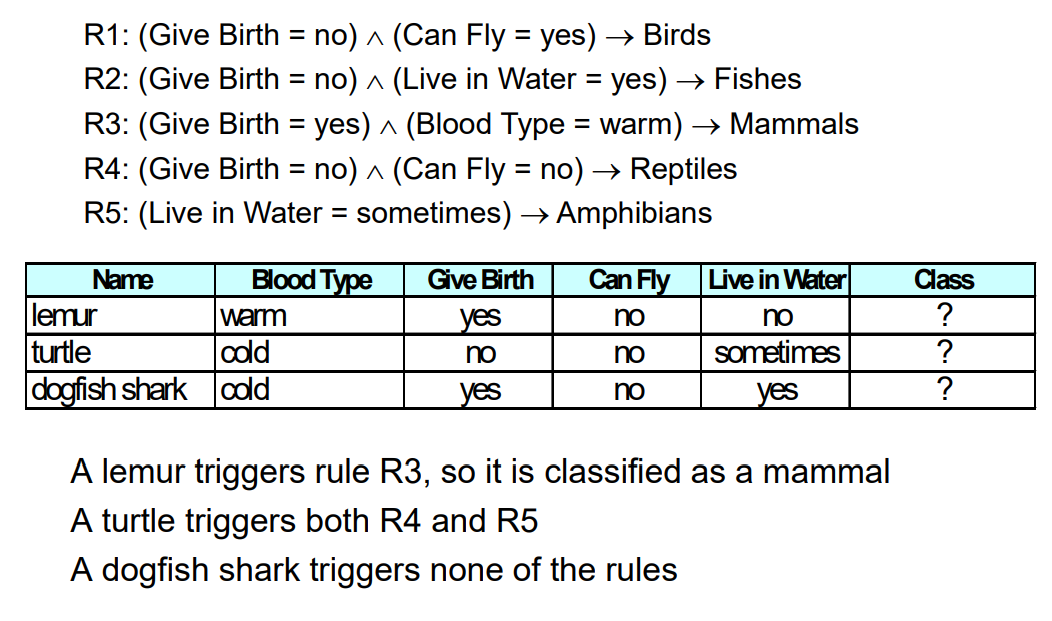
R1: (Give Birth = no) ^ (Can Fly = yes) => Birds

R2: (Give Birth = no) ^ (Live in Water = yes) => Fishes

R3: (Give Birth = yes) ^ (Blood Type = warm) => Mammals

R4: (Give Birth = no) ^ (Can Fly = no) => Reptiles

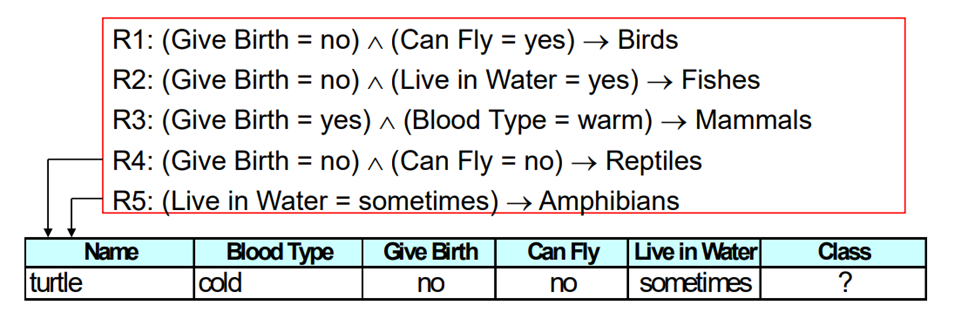
R5: (Live in Water = sometimes) => Amphibians



- Một con vượn cáo kích hoạt quy tắc R3, vì vậy nó được phân loại là động vật có vú

- Một con rùa kích hoạt cả R4 và R5

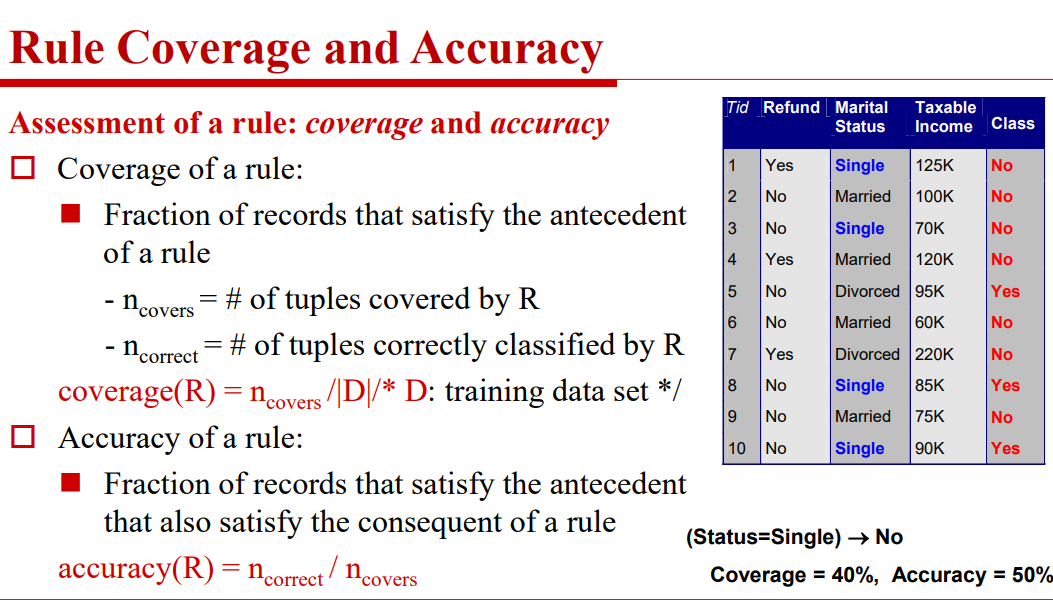
- Một con cá mập chó không kích hoạt bất kỳ quy tắc nào



- R4: (Give Birth = no) AND (Can Fly = no) → Reptiles (Bò sát) → thỏa mãn điều kiện này.

- R5: (Live in Water = sometimes) → Amphibians (Lưỡng cư) → cũng thỏa mãn điều kiện này.

* Ví dụ :Đánh giá quy tắc:



(Status=Single) => No

Coverage = 40%, Accuracy = 50%

Giải thích:

- Coverage = 40%: Quy tắc này áp dụng được cho 40% dữ liệu (những người có tình trạng hôn nhân là "Single").

- Accuracy = 50%: Trong số những người có tình trạng "Single", chỉ 50% được dự đoán chính xác là "No".

1. Tài liệu tham khảo:

* <https://jcsites.juniata.edu/faculty/rhodes/ml/rulebasedClass.htm>