

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**ﻣKHOA CÔNG NGHỆ THÔNG TINﻣ**

****

**BÀI TẬP LỚN**

**MÔN: HỆ CHUYÊN GIA**

**ĐỀ TÀI:** XÂY DỰNG HỆ CHUYÊN GIA TƯ VẤN MUA MÁY TÍNH

Giáo viên hướng dẫn: Lê Thị Thủy

Lớp: Hệ thống thông tin 1 – K11

**Nhóm 11**

*Thành viên của nhóm*

1. Lê Thị Hà
2. Chu Trọng Sơn
3. Nguyễn Xuân Trường

**Hà Nội – 2019**

**Mục Lục**

[Chương 1: Tổng quan về hệ chuyên gia và ứng dụng 3](#_Toc24746338)

[1.1. Hệ chuyên gia là gì? 3](#_Toc24746339)

[1.2. Đặc trưng hệ chuyên gia 5](#_Toc24746340)

[1.3. Lý do xây dựng hệ chuyên gia 6](#_Toc24746341)

[Chương 2: Giới thiệu về thuật toán cây quyết định 7](#_Toc24746342)

[2.1. Cây quyết định là gì? 7](#_Toc24746343)

[2.2. Thuật toán tạo cây quyết định 9](#_Toc24746344)

[2.2.1. Ý tưởng chính của thuật toán 9](#_Toc24746345)

[2.2.2. Thuật toán ID3 9](#_Toc24746346)

[2.3. Minh họa thuật toán ID3 11](#_Toc24746347)

[2.3.1. Dữ liệu mẫu và tính trên giấy 13](#_Toc24746348)

[2.3.2. Tập luật thu được 18](#_Toc24746349)

[Chương 3: Xây dựng chương trình tư vấn mua máy tính 20](#_Toc24746350)

[3.1. Thiết kế cơ sở dữ liệu 20](#_Toc24746351)

[3.2. Thiết kế giao diện chương trình 20](#_Toc24746352)

# Tổng quan về hệ chuyên gia và ứng dụng

## Hệ chuyên gia là gì?

Giáo sư Edward Feigenbaum của trường đại học STANFORD, một trong những chuyên gia đầu nghành về hệ chuyên gia, đã cho rằng:

Hệ chuyên gia là một hệ thống chương trình máy tính chứa các tri thức và quá trình suy diễn về một lĩnh vực cụ thể nào đó để giải quyết các bài toán khó nào đó mà đòi hỏi sự uyên bác của các chuyên gia trong ngành.

Hệ chuyên gia có các thành phần cơ bản sau:

1. Bộ giao diện người- máy.

2. Mô tơ suy diễn.

3. Cơ sở tri thức.

4. Bộ giải thích.

5. Bộ tiếp nhận tri thức.

6. Bộ nhớ làm việc.

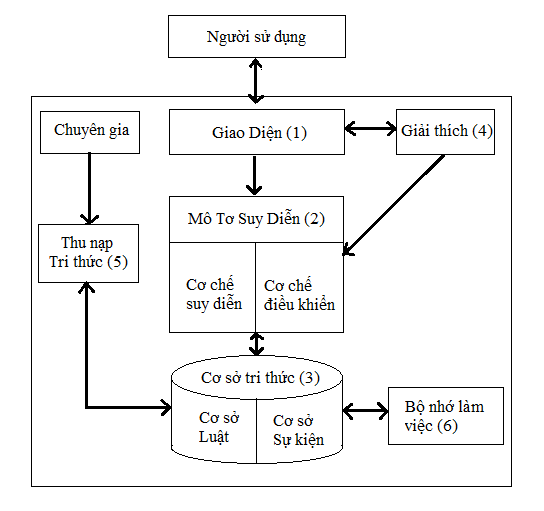
*Bộ giao diện người- máy (User Interface):* Thực hiện giao tiếp giữa HCG và người sử dụng. Bộ này nhận thông tin từ người sử dụng và đưa ra các câu trả lời, các lời khuyên, các giải thích về lĩnh vực đó.

*Mô tơ suy diễn (Interface Engine):* HCG mô hình hóa cách lập luận của con người với modul động cơ suy diễn. HCG chứa động cơ suy diễn để tiến hàng các suy diễn nhằm tạo ra các tri thức mới dựa trên sự kiện, tri thức trong vùng nhứ làm việc và trong cơ sở tri thức. Hai kiểu suy diễn chính trong động cơ suy diễn là suy diễn tiến và suy diễn lùi.

*Cơ sở tri thức (knowledge Base):* lưu trữ, biểu diễn các tri thức mà hệ đảm nhận, làm cơ sở cho các hoạt động của hệ. CSTT bao gồm cơ sở sự kiện (facts) và cơ sở luật (rules).

*Bộ giải thích (Explantion System):*  Trả lời câu hỏi là tại sao (why) và bằng cách nào (how) khi có yêu cầu của người sử dụng. Câu hỏi WHY nhằm mục đích cung cấp các lý lẽ để thuyết phục người sử dụng đi theo con đường suy diễn của hệ chuyên gia. Câu hỏi HOW nhằm cung cấp các giải thích về con đường mà hệ chuyên gia sử dụng để mang lại hiệu quả.

*Bộ tiếp nhận tri thức (knowledge Editor):* Làm nhiệm vụ nhận tri thức từ chuyên gia con người (human expert), từ kỹ sư xử lý tri thức và người sử dụng thông qua các yêu cầu và lưu trữ vào CSTT.



Hình 1.2. Các thành phần của hệ chuyên gia.

*Vùng nhớ làm việc (Working Memory):* Chứa các sự kiện liên quan được phát hiện trong quá trình đưa ra kết luận. Bộ nhớ làm việc tương đương với bộ nhớ ngắn hạn (Short- Term Memory) trong mô hình giải vấn đề của con người.

HCG = (1) + (2) + (3) + (4) + (5) + (6), trong đó (2) và (3) là bắt buộc phải có khi xây dựng HCG.

Mỗi hệ chuyên gia chỉ đặc trưng cho một lĩnh vực vấn đề (poblem domain) mà không phải cho bất cứ một lĩnh vực vấn đề nào.

Tri thức chuyên gia để giải quyết một vấn đề đặc trưng được gọi là lĩnh vực tri thức (knowledge domain).

Ví dụ: Hệ chuyên gia về lĩnh vực y học để chuẩn đoán bệnh sẽ có nhiều tri thức về một số triệu chứng của bệnh, lĩnh vực tri thức y học bao gồm các căn bệnh, triệu chứng và chữa trị.

*Chú ý*: rằng lĩnh vực tri thức hoàn toàn nằm trong lĩnh vực vấn đề. Phần bên ngoài lĩnh vực tri thức nói lên rằng không phải là tri thức cho tất cả vấn đề.

## Đặc trưng hệ chuyên gia

Tách tri thức khỏi điều khiển: Cơ sở tri thức và môtơ suy diễn độc lập với nhau. Điều này giúp việc phát triển, bảo trì hệ chuyên gia được thuận lợi hơn. Nghĩa là ta có thể bổ sung hay loại các sự kiện, các luật mà không làm ảnh hưởng tới động cơ suy diễn. Khi thay đổi điều khiển ta chỉ cần chỉnh sửa thuật toán trong động cơ suy diễn.

Có tri thức chuyên gia: Một đặc tính quan trọng của tri thức được sử dụng trong hệ chuyên gia đó là tri thức của chuyên gia. Tri thức chuyên gia được thu nhận và mã hóa trong hệ chuyên gia. Nó bao gồm tri thức lĩnh vực và kỹ năng giải quyết bài toán của chuyên gia.

Tập trung nguồn chuyên gia: Hầu hết các chuyên gia có kỹ năng giỏi giải quyết bài toán thuộc lĩnh vực của họ. Hệ chuyên gia cũng tương tự như vậy, nó chỉ tinh thông các vấn đề được huấn luyện, còn các vấn đề bên ngoài nó khó có khả năng giải quyết được.

Lập luận dựa trên các ký hiệu: Hệ chuyên gia biểu diễn tri thức ở dạng ký hiệu. Ta có thể sử dụng ký hiệu để biểu diễn nhiều dạng tri thức khác nhau. Hệ chuyên gia giải bài toán bằng cách xử lý các ký hiệu thay vì là xử lý số. Nói chung chương trình thông thường xử lý dữ liệu còn hệ chuyên gia xử lý tri thức.

Lập luận may rủi (reasons heuristically): Các chuyên gia tinh thông trong việc sử dụng các kinh nghiệm của họ để giải quyết bài toán đang xét một cách hiệu quả. Bằng các kinh nghiệm mà họ hiểu vấn đề qua thực tế và giữ nó dưới dạng may rủi. Hầu hết TTNT thủa ban đầu đều áp dụng các kỹ thuật tìm kiếm may rủi khi giải quyết vấn đề.

Khả năng giải quyết vấn đề bị hạn chế: Nếu không có chuyên gia để giải vấn đề thì ta cũng khó có thể hy vọng hệ chuyên gia giải tốt hơn. Nếu vấn đề quá mới hoặc thay đổi quá nhanh thì thực sự không có hệ chuyên gia nào có thể giải được. Ta chỉ nên xây dựng hệ chuyên gia để giải những bài toán mà chuyên gia giải được.

Độ phức tạp của bài toán: Các bài toán nên có độ phức tạp lập luận không quá dễ và cũng không quá khó. Vấn đề không quá phức tạp đến mức gây ra tình trạng không quản lý nổi ở tầm của chuyên gia. Nếu vấn đề có độ phức tạp lớn thì cố gắng chia thành các bài toán nhỏ hơn, mỗi bài toán con này ta có thể giải quyết bằng một hệ chuyên gia.

Chấp nhận sai lầm: Hệ chuyên gia giải quyết vấn đề như chuyên gia nên có thể mắc sai lầm, khi đó ta thấy rằng chương trình truyền thống có ưu thế hơn hệ chuyên gia. Tuy nhiên so sánh tổng thể người ta thấy rằng trong một số trường hợp hệ chuyên gia có lời giải sai lầm mang tính người hơn do xử lý các thông tin không chính xác, thậm chí mâu thuẫn nhau.

## Lý do xây dựng hệ chuyên gia

Có nhiều vấn đề xảy ra mà chuyên gia cũng như chương trình truyền thống không thể giải quyết được, hoặc có thể giải quyết được thì cũng tốn kém cả về tiền bạc, thời gian và sức lực. Khi đó, người ta thường chú ý đến xây dựng các HCG phục vụ cho từng mục đích khác nhau.

HCG có thể hoạt động như một chuyên gia trong việc truy tìm thông tin từ nhiều nguồn, từ nhiều chuyên gia.

HCG giữ lâu dài các tri thức chuyên gia ngay cả khi chuyên gia mất đi.

HCG cho kết quả bền vững, không bị cảm tính và thất thường như con người.

Tốc độ của HCG tỏ ra ưu việt, nhất là khi xử lý nhiều vấn đề cùng lúc.

Chi phí cho chuyên gia cao và có xu hướng tăng lên trong khi giá của HCG giảm xuống.

*Một số lý do HCG được phát triển để thay thế các chuyên gia đó là:*

* + Hiệu quả cao (High performance): Khả năng trả lời với mức độ tinh thông bằng hoặc cao hơn so với chuyên gia cùng lĩnh vực.
  + Thời gian trả lời thỏa đáng (Adequate response time): Thời gian trả lời hợp lý, bằng hoặc nhanh hơn so với chuyên gia để đi đến cùng một quyết định. Hệ chuyên gia là một hệ thống thời gian thực (real time System).
  + Độ tin cậy cao (high reliability): Không xảy ra sự cố hay giảm độ tin cậy khi sử dụng.
  + Dễ hiểu (understandable): HCG giải thích các bước suy luận một cách dễ hiểu và nhất quán, không giống như cách trả lời bí ẩn của các họp đen (black box).
  + HCG thay thế chuyên gia tại những nơi xa, nguy hiểm.
  + Việc tự động hóa công việc trong dây chuyền cần đến HCG mà con người không thể đáp ứng được.
  + HCG không những trợ giúp người bình thường mà còn hỗ trợ các chuyên gia.
  + Dùng lại các tri thức chuyên gia khi chuyên gia không còn nhớ được.

# Giới thiệu về thuật toán cây quyết định

## Cây quyết định là gì?

Việc học trên cây quyết định dựa trên dữ liệu huấn luyện có dạng bảng, trong đó các cột là thuộc tính điều kiện và thuộc tính mục tiêu(quyết định), mỗi dòng sẽ được dán nhãn là giá trị của thuộc tính mục tiêu. Chẳng hạn với bảng dữ liệu (Bảng 3.1) các thuộc tính “Age” và “Salary” là thuộc tính điều kiện còn thuộc tính “Class” là thuộc tính mục tiêu.

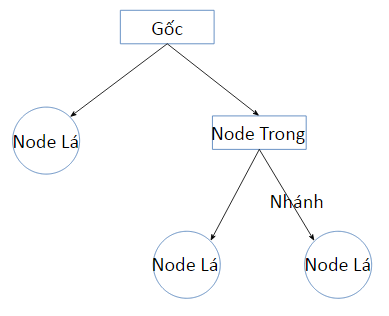
|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Age** | **Salary** | **Class** |
| 1 | 30 | 65 | Good |
| 2 | 23 | 15 | Bad |
| 3 | 40 | 75 | Good |
| 4 | 55 | 40 | Bad |
| 5 | 55 | 100 | Good |
| 6 | 45 | 60 | Good |

**Bảng 3.1**. *Dữ liệu huấn luyện*

Cây quyết định là một công cụ khá phổ biến trong nhiều dạng ứng dụng, với cơ chế rút trích các dạng luật if- then dựa trên tập liệu mẫu. Sau đây là ***định nghĩa cây quyết định.***

Cây quyết định là dạng đặc biệt có cấu trúc cây nhằm hỗ trợ cho tiến trình ra quyết định. Cây gồm các thành phần:

* Nút trong: Biểu diễn thuộc tính.
* Nhánh: Biểu diễn giá trị cụ thể của thuộc tính đó.
* Nút lá: Đại diện cho giá trị dự đoán của thuộc tính mục tiêu.



***Hình 3.1.:*** *Mô tả cây quyết định*

Good

Bad

Bad

Good

H<=35

h>35

H<=40

h>40

H<=50

h>50

***Hình 3.2.*** *Cây quyết định tương ứng với bảng 4.1*

Dựa trên cây này ta có thể đưa ra được quyết định.

Chẳng hạn với mẫu x=(30,30,?) sẽ được gán nhãn là “Bad”.

## Thuật toán tạo cây quyết định

### Ý tưởng chính của thuật toán

* Phương pháp “tham lam” (greedy).
* Phân chia tập mẫu dựa trên thuộc tính cho kết quả tối ưu hóa tiêu chuẩn.
* Điều kiện dừng:
* Tất cả các mẫu rơi vào một nút thuộc về cùng một lớp.
* Không còn thuộc tính nào có thể dùng để phân chia mẫu nữa.
* Không còn lại mẫu nào tại nút.
* Cách phân chia mẫu:
* Tiêu chuẩn phân chia: tạo ra các nhóm sao cho một lớp chiếm ưu thế trong từng nhóm.
* Thuộc tính được chọn là thuộc tính cho độ đo tốt nhất, có lợi cho quá trình phân lớp.
* Độ đo để đánh giá chất lượng phân chia là độ đo sự đồng nhất:
* Entropy (Information Gain)
* InformationGainRatio
* GiniIndex

*Một số thuật toán xây dựng cây quyết định:*

* Cart
* **ID3**, C4.5
* SLIQ, SPRINT

### Thuật toán ID3

Với bảng dữ liệu huấn luyện có nhiều thuộc tính điều kiện. Vấn đề đặt ra là lựa chọn thuộc tính nào làm gốc của cây. Để lựa chọn thuộc tính làm gốc của cây ta đưa ra tiêu chuẩn để đánh giá, một trong các tiêu chuẩn đó là độ lợi thông tin(information Gain) của thuộc tính. Thuộc tính có độ lợi thông tin lớn nhất sẽ được chọn làm gốc của cây.

Độ lợi thông tin của thuộc tính A, ký hiệu là Gain(S, A) và được xác định qua công thức:

**Gain(S, A)=Entropy(S) -∑((|Sv | / |S|)\* Entropy(Sv) )**

**Thuật toán ID3:**

Cho X là tập vectơ đặc trưng( không gian đặc trưng):

X={x1, x2,...,xn}

C là tập các giá trị phân lớp

C: X ─> C thực hiện gán nhãn cho các phần tử trong X

D= { D = {(x1, c(x1)), (x2, c(x2)),…, (xn, c(xn))} ⊆ X × C}

Nhiệm vụ: Dựa trên D, hãy xây dựng cây quyết định T xấp xỉ c.

**ID3( D, Atributes, Target)**

{

* Tạo nút t cho cây.
* Nếu tất cả các mẫu trong D đều thuộc lớp dương thì trả về cây có nút gốc t và được dán nhãn “+”.
* Nếu tất cả các mẫu trong D đều thuộc lớp âm thì trả về cây có nút gốc t và được dán nhãn “-”.
* Đặt t là nhãn phổ biến nhất của thuộc tính mục tiêu trong D.
* Nếu Atributes rỗng thì trả về cây có nút gốc trỏ bởi t.
* Trường hợp khác:
* Đặt A\* là thuộc tính phân lớp tốt nhất trong D.
* Thuộc tính quyết định của t là A\*.
* Với mỗi giá trị a của A\*  thực hiện:
  + - Bổ sung nhánh mới dưới t ứng với A\* = “a”.
    - Đặt D\_a là tập con của D chứa các mẫu mà A\* = “a”.
    - Nếu D\_a rông thì nhánh mới này bổ sung nút lá với nhãn phổ biến nhất trong D.
    - Ngược lại, dưới nhánh mới này bổ sung cây con ID3( D\_a, Atributes – { A\*}, Target).
* Trả về t.

}

## Minh họa thuật toán ID3

Để minh họa cho thuật toán này, ta xét tập dữ liệu huấn luyện sau:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Age** | **Sex** | **Income** | **Configuration** | **Manufacturers** | **Os** | **Buy** |
| 1 | youth | male | High | high | SONY | Windows | Yes |
| 2 | senior | Male | Low | low | DELL | Windows | Yes |
| 3 | senior | female | High | high | APPLE | Mac OS | Yes |
| 4 | youth | Male | High | medium | DELL | Linux | No |
| 5 | middle-aged | Male | Medium | high | APPLE | Mac OS | Yes |
| 6 | youth | female | High | high | DELL | Windows | No |
| 7 | senior | Male | Low | medium | SAMSUNG | Linux | Yes |
| 8 | youth | Male | Low | low | HP | Linux | Yes |
| 9 | middle-aged | female | Low | high | APPLE | Mac OS | No |
| 10 | middle-aged | Male | Medium | medium | APPLE | Mac OS | Yes |
| 11 | youth | female | Medium | high | ASUS | Windows | No |
| 12 | youth | Male | High | low | ASUS | Windows | Yes |
| 13 | senior | Male | Medium | medium | HP | Windows | Yes |
| 14 | youth | female | Low | high | DELL | Windows | No |
| 15 | middle-aged | Male | low | medium | APPLE | Mac OS | No |
| 16 | senior | Male | Medium | low | DELL | Windows | Yes |
| 17 | senior | female | Low | medium | SONY | Windows | No |
| 18 | middle-aged | Male | High | low | HP | Linux | No |
| 19 | youth | Male | Medium | high | HP | Linux | No |
| 20 | senior | female | Medium | low | APPLE | Mac OS | Yes |
| 21 | senior | Female | low | medium | APPLE | Mac OS | No |
| 22 | youth | female | Medium | medium | ASUS | Windows | Yes |
| 23 | middle-aged | Male | Low | high | SAMSUNG | Windows | No |
| 24 | youth | female | High | medium | DELL | Windows | Yes |
| 25 | youth | Male | Low | high | ASUS | Linux | No |
| 26 | middle-aged | female | Medium | medium | APPLE | Mac OS | No |
| 27 | senior | Male | High | medium | SONY | Windows | Yes |
| 28 | senior | Male | Medium | high | APPLE | Mac OS | Yes |
| 29 | middle-aged | female | High | medium | APPLE | Mac OS | No |
| 30 | youth | Male | Medium | medium | APPLE | Mac OS | Yes |
| 31 | senior | female | Low | high | ASUS | Linux | Yes |
| 32 | middle-aged | Male | Low | low | APPLE | Mac OS | No |
| 33 | senior | Male | High | high | ASUS | Windows | Yes |
| 34 | senior | female | High | low | DELL | Windows | Yes |
| 35 | middle-aged | female | Medium | low | SONY | Windows | Yes |
| 36 | youth | Male | Medium | low | HP | Linux | Yes |
| 37 | middle-aged | Male | High | low | ASUS | Linux | No |
| 38 | senior | female | Medium | high | ASUS | Windows | No |
| 39 | youth | female | Low | medium | HP | Windows | Yes |
| 40 | middle-aged | female | Medium | high | APPLE | Mac OS | Yes |
| 41 | senior | Male | medium | high | SONY | Mac OS | No |
| 42 | youth | Male | low | medium | SAMSUNG | Linux | Yes |
| 43 | middle-aged | female | high | high | DELL | Windows | Yes |
| 44 | youth | female | medium | low | HP | Linux | No |
| 45 | middle-aged | Male | high | medium | DELL | Windows | Yes |
| 46 | middle-aged | Male | medium | low | HP | Windows | Yes |
| 47 | senior | female | high | medium | HP | Windows | No |

### Dữ liệu mẫu và tính trên giấy

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Age** | **Sex** | **Income** | **Configuration** | **Manufacturers** | **Os** | **Buy** |
| 1 | youth | Male | high | high | SONY | Windows | Yes |
| 2 | senior | Male | low | low | DELL | Windows | Yes |
| 3 | senior | Female | high | high | APPLE | Mac OS | Yes |
| 4 | youth | Male | high | medium | DELL | Linux | No |
| 5 | middle-aged | Male | medium | high | APPLE | Mac OS | Yes |
| 43 | middle-aged | Female | high | high | DELL | Windows | Yes |
| 44 | youth | Female | medium | low | HP | Linux | No |
| 45 | middle-aged | Male | high | medium | DELL | Windows | Yes |
| 46 | middle-aged | Male | medium | low | HP | Windows | Yes |
| 47 | senior | Female | high | medium | HP | Windows | No |

**Bảng 3.2.** Số liệu quan sát về người dùng mua máy tính

Entropy(S)= \* log2 \* log2 = **0.881**

* **Phân hoạch lần thứ 1:**

*Xét thuộc tính* ***Age:***

Entropy(Syouth) = \* log2 \* log2 = 0.918

Entropy(Ssenior) = \* log2 \* log2 = 0.918

Entropy(S middle-aged) = \* log2 = 0

=>Gain(S,Age)= Entropy(S)\* Entropy(Syouth)\* Entropy(Ssenior)

\* Entropy(S middle-aged)

= **0.330**

*Xét thuộc tính* **Sex*:***

Entropy(SMale) = \* log2 \* log2 = 0.650

Entropy(SFemale) = \* log2 \* log2 = 1

=>Gain(S,Sex)= Entropy(S)\* Entropy(SMale) \* Entropy(SFemale) = **0.091**

*Xét thuộc tính* **Income*:***

Entropy(S high) = \* log2 \* log2 = 0.918

Entropy(Smedium) = \* log2 \* log2 = 0.918

Entropy(Slow) = \* log2 = 0

=>Gain(S,Income)= Entropy(S)\* Entropy(S high) \* Entropy(Smedium)

\* Entropy(Slow)

= **0.055**

*Xét thuộc tính* **Configuration*:***

Entropy(Shigh) = \* log2 = 0

Entropy(Smedium) = \* log2 \* log2 = 0.918

Entropy(Slow) = \* log2 \* log2 = 0.918

=>Gain(S,Configuration)= Entropy(S)\* Entropy(Shigh) \* Entropy(Smedium)

\* Entropy(Slow)

= **0.330**

*Xét thuộc tính* **Manufacturers*:***

Entropy(SSONY) = \* log2 = 0

Entropy(SDELL) = \* log2 \* log2 = 0.811

Entropy(SAPPLE) = \* log2 = 0

Entropy(SHP) = \* log2 \* log2 = 0.918

=>Gain(S,Manufacturers)= Entropy(S)\* Entropy(SSONY) \* Entropy(SDELL)

\* Entropy(SAPPLE)\* Entropy(SHP)

= **0.281**

*Xét thuộc tính* **Os*:***

Entropy(SWindows) = \* log2 \* log2 = 0.650

Entropy(SMac OS) = \* log2 = 0

Entropy(SLinux) = \* log2 = 0

=>Gain(S,OS)= Entropy(S)\* Entropy(SWindows)\* Entropy(SMac OS)

\* Entropy(SLinux)

= **0.491**

Trong các thuộc tính điều kiện trên ta thấy “OS” có độ lợi thông tin lớn nhất nên thuộc tính này sẽ được chọn làm gốc của cây.

* **Phân hoạch lần thứ 2:**

*Xét thuộc tính* ***Age:***

Entropy(Syouth) = \* log2 = 0

Entropy(Ssenior) = \* log2 \* log2 = 1

Entropy(S middle-aged) = \* log2 = 0

=>Gain(SWindows,Age)= Entropy(SWindows)\* Entropy(Syouth)

\* Entropy(Ssenior)\* Entropy(S middle-aged)

=0.316

*Xét thuộc tính* **Sex*:***

Entropy(SMale) = \* log2 = 0

Entropy(SFemale) = \* log2 \* log2 = 1

=>Gain(SWindows,Sex)= Entropy(SWindows)\* Entropy(SMale)

\* Entropy(SFemale)

= 0.316

*Xét thuộc tính* **Income*:***

Entropy(S high) = \* log2 \* log2 = 0.811

Entropy(Smedium) = \* log2 = 0

Entropy(Slow) = \* log2 = 0

=>Gain(SWindows,Income)= Entropy(SWindows)\* Entropy(S high)

\* Entropy(Smedium)\* Entropy(Slow)

=0.109

*Xét thuộc tính* **Configuration*:***

Entropy(Shigh) = \* log2 = 0

Entropy(Smedium) = \* log2 \* log2 = 1

Entropy(Slow) = \* log2 = 0

=>Gain(SWindows,Configuration)= Entropy(SWindows)\* Entropy(Shigh)

\* Entropy(Smedium)\* Entropy(Slow)

=0.316

*Xét thuộc tính* **Manufacturers*:***

Entropy(SSONY) = \* log2 = 0

Entropy(SDELL) = \* log2 = 0

Entropy(SHP) = \* log2 \* log2 = 1

=>Gain(SWindows,Manufacturers)= Entropy(SWindows)\* Entropy(SSONY)

\* Entropy(SDELL)\* Entropy(SHP)

=0.316

Trong các thuộc tính điều kiện trên ta thấy **Age, Sex, Configuration, Manufacturers** có độ lợi thông tin lớn nhất nên chọn thuộc tính **Sex** được chọn làm gốc của cây.

* **Phân hoạch lần thứ 3:**

*Xét thuộc tính* ***Age:***

Entropy(Smiddle-aged) = \* log2 = 0

Entropy(Ssenior) = 0 \* log2 = 0

=>Gain(SFemale,Age)= Entropy(SFemale)\* Entropy(Ssenior)

\* Entropy(S middle-aged)

= 1

*Xét thuộc tính* **Income*:***

Entropy(S high) = = \* log2 \* log2 = 1

=>Gain(SFemale,Income)= Entropy(SFemale)\* Entropy(S high)

=0

*Xét thuộc tính* **Configuration*:***

Entropy(Shigh) = \* log2 = 0

Entropy(Smedium) = 0 \* log2 = 0

=>Gain(SFemale,Configuration)= Entropy(SFemale)\* Entropy(Shigh)

\* Entropy(Smedium)

= 1

*Xét thuộc tính* **Manufacturers*:***

Entropy(SDELL) = \* log2 = 0

Entropy(SHP) = 0 \* log2 = 0

=>Gain(SFemale,Manufacturers)= Entropy(SFemale)

\* Entropy(SDELL)\* Entropy(SHP)

= 1

Trong các thuộc tính điều kiện trên ta thấy **Age, Configuration, Manufacturers** có độ lợi thông tin lớn nhất nên chọn thuộc tính **Age** được chọn làm gốc của cây.

Windows

Mac OS

Linux

Yes

No

Male

Female

Yes

middle-aged

senior

Yes

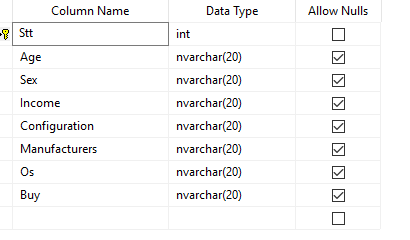
No

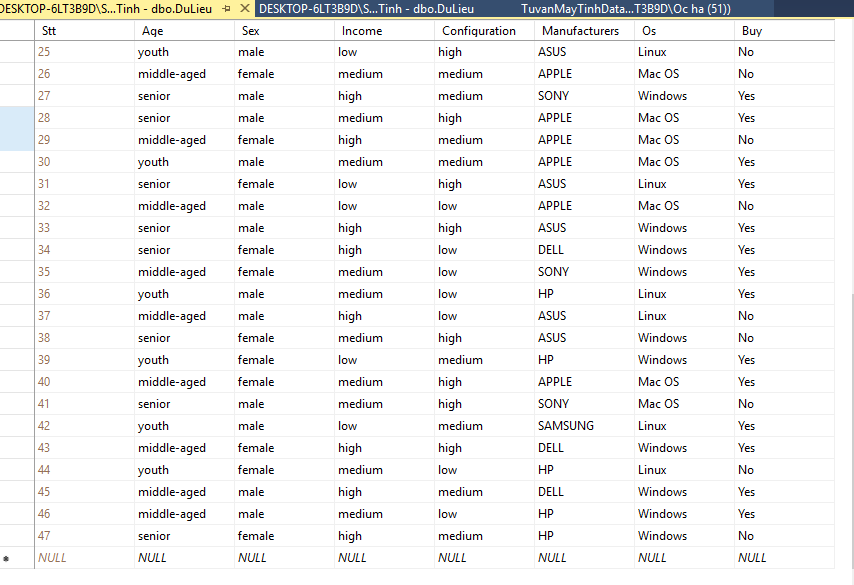
### Tập luật thu được

* Rule 1: IF OS = Linux THEN Buy = No
* Rule 2: IF OS = MacOS THEN Buy = Yes
* Rule 3: IF OS = Windows AND Sex = Male THEN Buy = Yes
* Rule 4: IF OS = Windows AND Sex = Female AND Age =senior THEN Buy = No
* Rule 5: IF OS = Windows AND Sex = Female AND Age = middle-aged THEN Buy = Yes

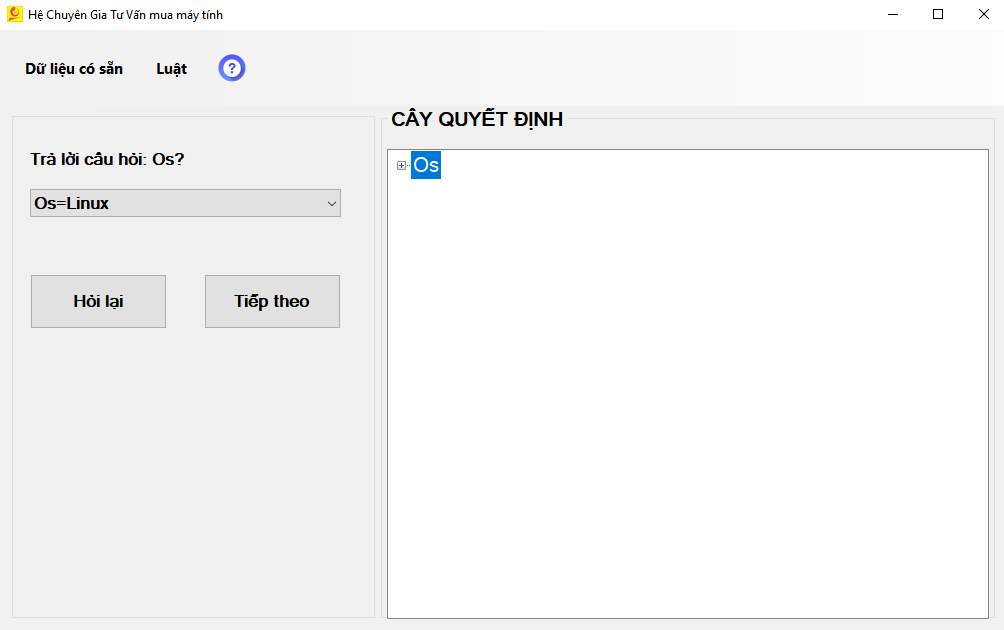
# Xây dựng chương trình tư vấn mua máy tính

## Thiết kế cơ sở dữ liệu

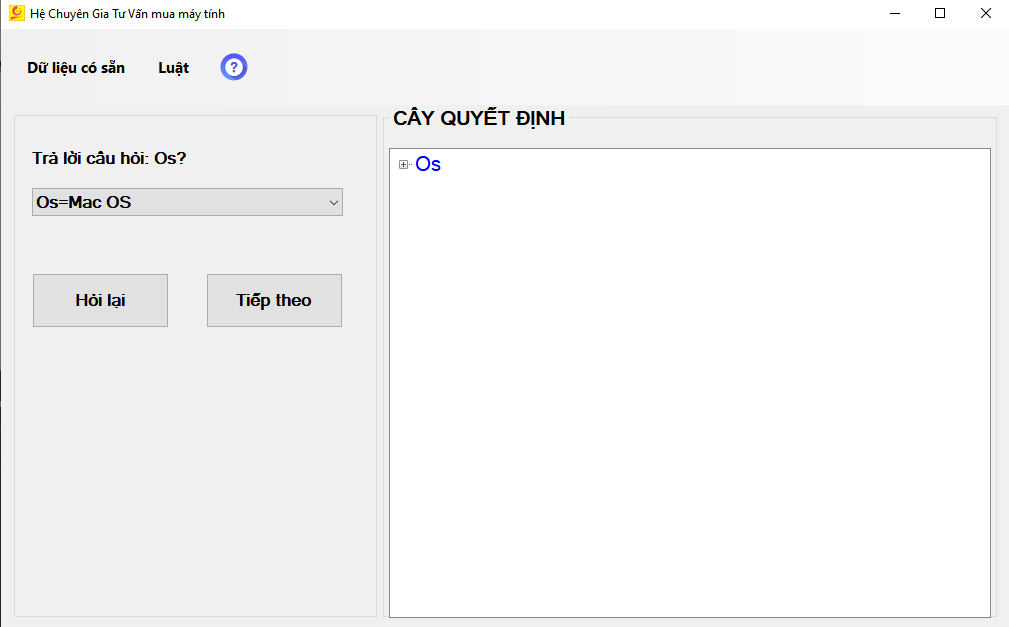




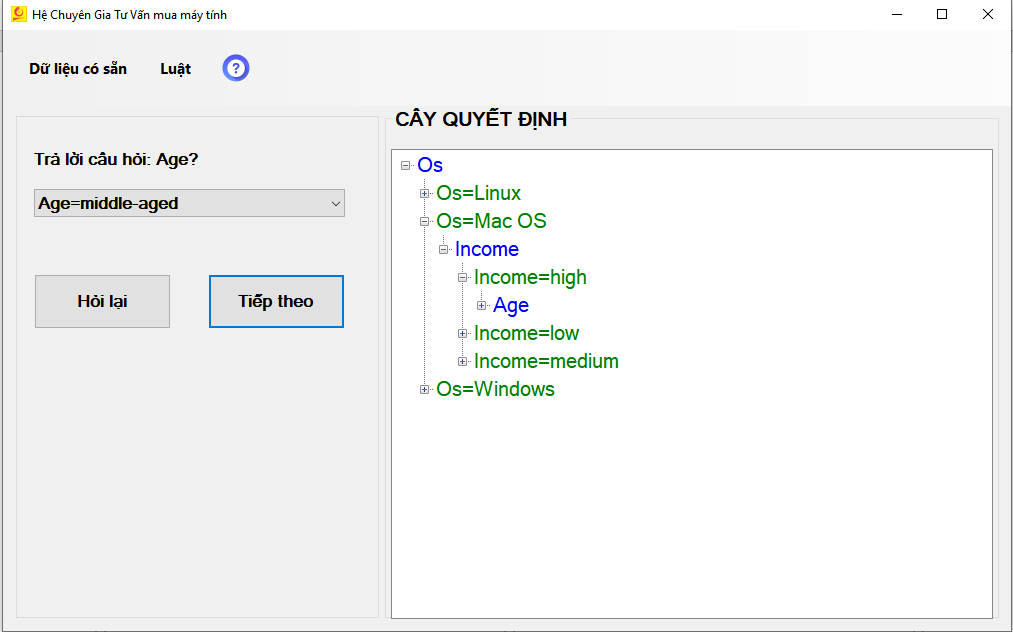
## Thiết kế giao diện chương trình



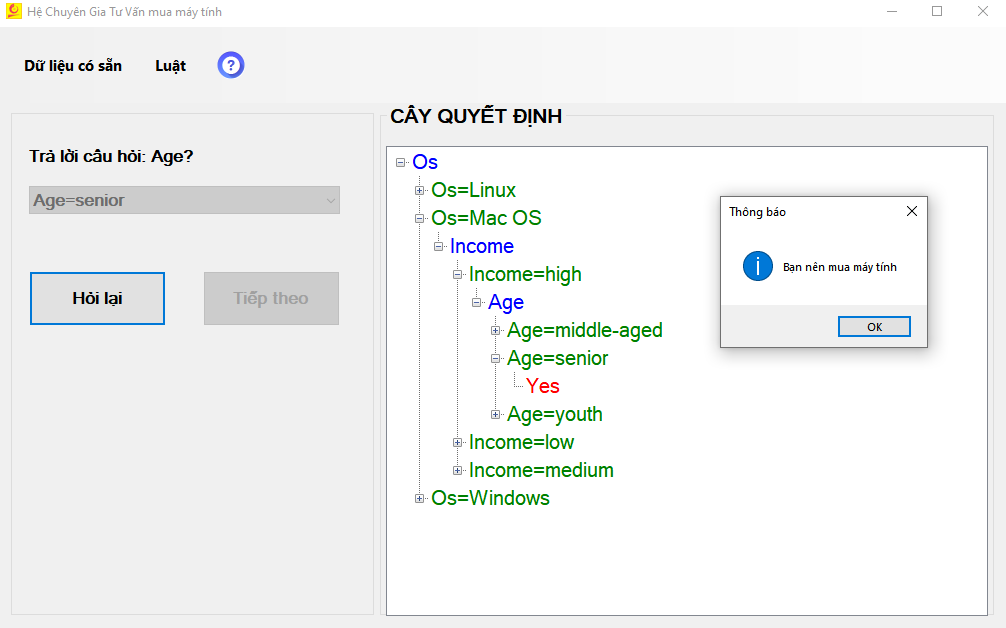
Hình 3.2‑1 Giao diện chương trình sau khi khởi tạo cây quyết định xong



Hình 3.2‑2 Ví dụ chọn nhánh OS=Mac OS



Hình 3.2‑3 Trong nhánh Mac OS lại chọn Income=high



Hình 3.2‑4 Tiếp tục chương trình lại hỏi Age, chọn Senior và chúng ta có được câu trả lời